# Informationsextraktion aus Stellenanzeigen im Internet

**Sandra Bsiri & Michaela Geierhos**
Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilians-Universität München
D-80538 München, Deutschland
{sandra.bsiri|michaela.geierhos}@cis.uni-muenchen.de

## Abstract

Dieser Beitrag beschäftigt sich mit der Informationsextraktion aus Stellenanzeigen im französischsprachigen Web. Ziel dieser Arbeit ist es, unstrukturierte Dokumente in Repräsentationsvektoren anhand lokaler Grammatiken zu transformieren. Auf diese Weise wird es möglich, den Stellenmarkt für Jobsuchmaschinen transparenter zu gestalten, indem nur auf dem Inhalt der Anzeige in Form von Darstellungsvektoren anstatt auf unübersichtlichem Fließtext gesucht werden muss.

## 1 Einführung

Das Internet hat sich dank seiner hohen Verbreitungseffizienz zum zentralen Medium des Stellenmarktes entwickelt [Fondeur *et al.*, 2005]. Gemäß einer Untersuchung, die der Personalvermittler *Kelly Services*[1] im ersten Quartal 2006 mit 19.000 beteiligten Personen aus 12 europäischen Ländern durchgeführt hat, ist das Internet inzwischen das an erster Stelle genutzte Kommunikationsmittel auf der Suche nach einem Arbeitsplatz. Die Studie kommt unter anderem zu dem Schluss, dass das Wachstumspotential des Internets bei der Arbeitsvermittlung keineswegs ausgeschöpft ist, was eine weitere, von der *Focus RH Gruppe* durchgeführte Studie [Focus RH, 2006] über die 500 wichtigsten Internet-Jobbörsen in Frankreich bestätigt.

Das Spektrum konkurrierender Stellenbörsen im Netz hat allerdings zu einer hohen Redundanz der Daten und eingeschränkter Transparenz geführt. So wird ein Großteil der auf dem Arbeitsmarkt verfügbaren freien Stellen über andere Kommunikationswege wie Firmenwebseiten, spezialisierte Diskussionsforen oder auch Kleinanzeigenverzeichnisse [Fondeur *et al.*, 2005] veröffentlicht. Dabei wird eine erhebliche Anzahl an Stellenangeboten unter heterogenen Formaten gleichzeitig auf mehreren Jobbörsen publiziert. Außerdem erschwert die weite Streuung der Stellenanzeigen im Internet ihre Zugänglichkeit für Web-Suchmaschinen und somit für den interessierten Bewerber.

In den letzten Jahren lag der Schwerpunkt bei vielen Online-Anbietern vorwiegend auf der Quantität der zur Verfügung gestellten Daten (der publizierten Stellenanzeigen bzw. der in den Datenbanken gespeicherten Lebensläufe) [2], wobei die Qualität der Suchergebnisse teilweise aus dem Blick verloren ging. Jedoch beschränken sich Jobsuchmaschinen dieses Typs in der Transformationsphase der Suchanfrage auf die traditionellen, rein statistischen Indexierungs- und Suchmethoden [Glöggler, 2003; Kowalski, 1997], anstatt den Vorteil der bereits semistrukturierten Daten [Ferber, 2003] zu nutzen und somit präzisere Suchergebnisse zu ermitteln. Deshalb sollte es das Ziel einer jeden Jobsuchmaschine sein, eine gewisse Transparenz in den Arbeitsmarkt zu bringen und die Fülle an offenen Stellenanzeigen virtuell zu vereinigen. Somit sollten die Ergebnisse *einer* spezifischen Suche alle ausschließlich relevanten Arbeitsangebote enthalten. Dabei wäre es wünschenswert, dass dies über eine einzige graphische Oberfläche und in Echtzeit für alle aktuellen, freien, online-publizierten Stellen realisiert würde.

Ähnlich wie bei [Flury, 2005] wird in dieser Arbeit am Fall des frankophonen Stellenmarktes illustriert wie mit linguistisch basierten Methoden alle im Netz verfügbaren Stellenangebote erstmals auf einer zentralisierten Plattform gesammelt und gefiltert werden können. Ein System zur automatischen Erkennung und Klassifikation von Firmen-Homepages wurde implementiert, um eine Firmendatenbank zu erstellen und auf aktuellem Stand zu halten. Basierend auf diesem Verzeichnis werden die HTML-Strukturen auf Ankertexte durchsucht, die zu den offenen Stellen führen. Lokale Grammatiken [Gross, 1993; 1997; Geierhos, 2006] und elektronische Lexika [Courtois *et al.*, 1990; 1997] wurden ausgearbeitet, um in einer zweiten Phase jene Informationen zu extrahieren, die für die Umwandlung der reinen Textform der Stellenangebote in ein semantisch strukturiertes Dokument notwendig sind. Über diese linguistische Analyse der einzelnen Einträge wird die Datenbank automatisch gefüllt und höchste Selektivität sowie Benutzerfreundlichkeit bei Suchanfragen gewährleistet.

Zur Realisierung einer verbesserten Jobsuchmaschine wurde im Rahmen dieser Arbeit ein Multi-Level-System aufgebaut, das in all seinen Analyse- und Bearbeitungsphasen von linguistischen Theorien und besonders vom Konzept „lokaler Grammatiken“ getragen wird. Das gesamte System (siehe Abbildung 1) basiert auf zwei hier in Abschnitt 2 und 3 ausgearbeiteten, interagierenden Modulen (A und B) und erinnert auf den ersten Blick an den Aufbau einer gewöhnlichen Suchmaschine:

**A** Lokalisieren der Stellenanzeigen im Web

**B** Dokumentanalyse und Informationsextraktion

**C** Bearbeitung der Suchanfragen

Jedoch liegt der Schwerpunkt unserer Ausführungen verstärkt auf Teil B und der damit verbundenen Informationsextratkion aus Stellenanzeigen, sowie ihrer Trans-

[1] siehe auch http://management.journaldunet.com/repere/outils-recherche-emploi.shtml

[2] siehe auch http://edito.keljob.com/index.php?id=27#1595 http://edito.keljob.com/recruteurs/articles/1206/barometre-octobre.html

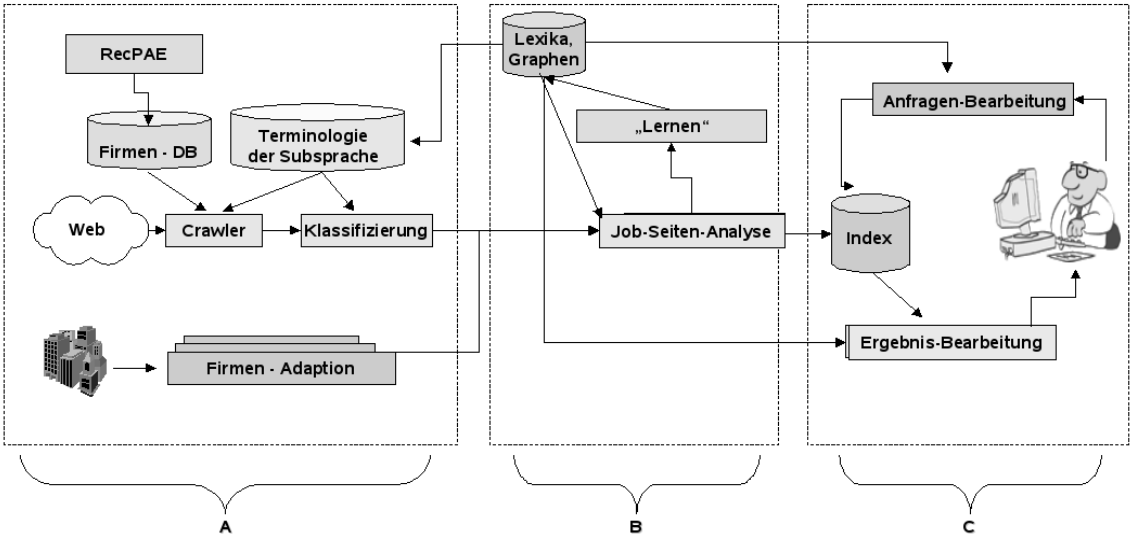


**Abbildung 1:** Übersicht zur Systemarchitektur

formation in Repräsentationsvektoren (Abschnitt 3), welche beim späteren Retrieval schnellere und qualitativ bessere Antworten auf Suchanfragen liefern sollen. So gibt Abschnitt 4 Einblick in die Ergebnisse der Systemevaluation und zeigt, wie vielversprechend die Werte für Precision (durchschnittlich 95.86 %) und Recall (durchschnittlich 92.98 %) bei der Erkennung anzeigenspezifischer Entitäten, wie z.B. Berufsbezeichnung, Arbeitsort etc. sind. Zuletzt werden noch mögliche Weiterentwicklungen und Anwendungen des hier beschriebenden Systems skizziert.

## 2 Lokalisieren von Stellenanzeigen im Web

Um die Idee einer zentralisierten Plattform für Stellenangebote zu verwirklichen, ist es notwendig, die im Internet verfügbaren Stellen automatisch zu finden. Zu diesem Zweck werden zwei unterschiedliche Strategien verfolgt: Einerseits werden bevorzugt Firmenwebseiten auf freie Stellen durchsucht, wofür ein System zur automatischen Erkennung dieser entwickelt wurde. Andererseits wird ein fokussierter Crawler eingesetzt, welcher mit Hilfe der speziellen Terminologie von Stellenanzeigen jede gecrawlte Website als solche identifiziert und dementsprechend klassifiziert.

### 2.1 Identifikation von Firmenwebsites

Die eben genannte Vorgehensweise – basierend auf einer Datenbank aus Internetadressen der jeweiligen Unternehmen – ermöglicht es, die auf den Firmenhomepages veröffentlichten Stellen zu finden. Dafür muss eine solche Datenbasis erst zur Verfügung gestellt und aufgrund der stetigen Bewegung auf dem Arbeitsmarkt immer aktuell gehalten werden. So ist es zwingend notwendig, jede URL, die als Webauftritt eines Unternehmens klassifiziert wurde, täglich aufzusuchen und auf ihre Erreichbarkeit zu testen. Des Weiteren müssen die dort veröffentlichten Stellenangebote im Vergleich mit ihrem Status des Vortages und anhand ihres Auschreibungszeitpunktes auf ihre Aktualität überprüft werden. Diese Update-Strategie erfordert ein großes Repertoire an Crawler-Seed-Links (URLs), welches aber mittels des neu entwickelten Systems zur automatischen Erkennung von Firmenwebsites[3] angelegt werden kann. Dabei handelt es sich um ein automatisches binäres Klassifikationssystem, welches für jede URL entscheidet, ob sie der Klasse „Organisation“ angehört oder nicht. Für die Klassifikation werden Merkmalsvektoren zur Entscheidungsfindung eingesetzt, welche sich bereits während der Lernphase bewährt haben, und nun eine Zuordnung zu insgesamt 10 vordefinierten Klassen ermöglichen.

Folgende Kriterien beeinflussen maßgeblich die Klassifikation der potentiellen Firmenwebsites:

- Orientierung an der HTML-Struktur des gesamten Dokumentes
- Auswertung der URLs, Meta-Informationen und des Titels
- Analyse und Einordnung der Ankertexte in vordefinierte semantische Klassen (siehe Tabelle 1)
- Identifikation des Firmennamens
- Berücksichtigung der Adresse, Telefonnummer, Handelsregisternummer, etc.
- Extraktion typischer Formulierungen und komplexer Terminologie

Unter Berücksichtigung von Flash-animierten Homepages konnten ca. 72 % der potiellen Firmenwebsites als solche korrekt identifiziert werden. Schließt man jedoch die Menge dieser Seiten aus, ist die Erkennungsrate deutlich höher (86 %) und stellt eine ausbaufähige, aber dennoch solide Basis für unser Vorhaben dar.

**Firmennamenerkennung**

Die automatische Erkennung von Organisationsnamen taucht relativ oft in der Literatur auf [Mallchok, 2004] und ist ein Teilbereich der automatischen Erkennung von Eigennamen (Named-Entity-Recognition). Es hat sich gezeigt, dass eine Liste von Organisationsnamen nicht ausreichend ist, um eine hohe Annotierungsquote zu erreichen,

[3] Die Firmenwebsites werden hier pro Webauftritt *(web site)* und nicht pro Webseite *(web page)* klassifiziert.

| | |
|---|---|
| **Carrière** *(Einstellungsmöglichkeiten)* | Nous recrutons *(Wir stellen ein)*<br>Nos offres d'emploi *(Unsere Stellenangebote)* |
| **Produits/Services** *(Produkte/Dienstleist.)* | Nos Produits *(Unsere Produkte)*<br>Accès à la boutique *(Zum Shop)* |
| **Contact** *(Kontaktinformation)* | Nous contacter *(Kontaktieren Sie uns)*<br>Pour venir nous voir *(Besuchen Sie uns)*<br>Nos coordonnées *(Unsere Kontaktdaten)* |
| **Société** *(Firmeninformationen)* | Notre Société *(Unser Unternehmen)*<br>Qui sommes nous? *(Wer wir sind)*<br>Entreprise *(Unternehmen)* |
| **Presse** | Communiqués de Presse *(Pressemitteilungen)*<br>La presse et nous *(Wir in der Presse)*<br>Presse infos *(Presseinformationen)*<br>media *(Medien)* |
| **Clients/Partenaires** *(Kunden/Partner)* | Nos Clients *(Unsere Kunden)*<br>Espace clientèles *(Kundenbereich)*<br>Nos partenaires *(Unsere Partner)*<br>Relation client *(Kundenbeziehung)* |

**Tabelle 1:** Ausgewählte Beispiele klassifizierender Ankertexte

da sehr viele Ambiguitäten bei Firmenbezeichnungen existieren. Allerdings sind alle bereits veröffentlichten Systeme gleichermaßen stark von den jeweiligen Trainingskorpora oder von der verwendeten Subsprache [Harris, 1968; 1988] abhängig. Um nun die Leistung dieses Systems der automatischen Klassifikation von Firmenwebseiten zu verbessern, wird zusätzlich zu den oben beschriebenen Eigenschaften der Name der Unternehmen extrahiert. Genau diese Information ist nötig, um als einer der Deskriptoren zu fungieren, um letztendlich die Klassifikationsentscheidung zu treffen.

Insgesamt wurden zwei verschiedene Methoden zur Extraktion der Firmennamen ausgearbeitet. Einerseits wird das linguistische Konzept der lokalen Grammatiken[4] angewandt, um die entsprechenden firmentypischen Kontexte zu beschreiben. Andererseits wurde ein Algorithmus entwickelt, der die Segmentierung des Domänennamens der zu analysierenden Website vornimmt.

Lokale Grammatiken [Gross, 1997] ermöglichen die Beschreibung eines lokalen Kontextes und schränken bestimmte lexikalische oder snytaktische Einheiten auf ein Fenster fester Größe ein. Dadurch ermöglichen sie es, Mehrdeutigkeit zu vermeiden bzw. einzugrenzen. In der Tat sollen die Auslöser – die einleitenden Kontexte – eines semantisch-syntaktischen Musters identifiziert, und mittels Bootstrapping [Gross, 1999] dessen Kontexte detailliert beschrieben werden.

In dieser Phase liegt der Schwerpunkt ganz auf der Sammlung externer Kontexte der Organisationsnamen, welche ebenfalls durch ein Bootstrapping-Verfahren mit Graphen [Senellart, 1998a; 1998b] ermittelt wurden und anschließend darin Verwendung fanden.

**Floskeln und Firmenterminologie**
Für die Erkennung und Extraktion von firmentypischen Kenngrößen, Standardredewendungen (Floskeln) oder unternehmensspezifischen Kontexten sind lokale Grammatiken ein adäquater Formalismus, um präzise und modular die rechten und linken Kontexte lexikalischer Einheiten zu beschreiben. Auf diese Weise kann die hohe syntaktische Variabilität bewältigt werden, was klare Vorteile gegenüber einer normalen Stringsuche bietet. Denn eine gewöhnliche Suche würde eine Liste aller gesammelten Floskeln mit dem Text der Website abgleichen, was dazu führen kann, dass bestimmte Ausdrücke mit minimalen morphosyntaktischen Variationen nicht mehr gefunden werden. Im Gegensatz dazu ist Bootstrapping mittels lokaler Grammatiken ein vielversprechender Ansatz, was folgendes Beispiel illustriert:

- Notre société , leader mondial sur le marché [...]
- Notre société est leader européen dans le secteur [...]
- Notre société leader dans son domaine [...]

Obwohl diese drei Phrasen sich auf lexikosyntaktischer Ebene stark unterscheiden, beschreiben sie einen ähnlichen Kontext des Wortes „leader". Doch mit Hilfe diverser statistischer Methoden, auf denen das System zur Extraktion von Konzepten basiert, können Ausdrücke dieser Gestalt aufgrund ihrer geringen Frequenz ignoriert werden, da trotz der syntaktischen Unterschiede bei der Paraphrasierung die Semantik erhalten bleibt. Jedoch ist es möglich diese Paraphrasen in einer lokalen Grammatik zusammenzufassen, wie sie beispielsweise in Abbildung 2 dargestellt wird.

Sobald eine Firmenhomepage als solche erkannt wurde, beginnt die Suche nach den Stellenangeboten, welche sich am HTML-Gerüst der Seite orientiert. Zu diesem Zweck wurde die Klasse der „Jobs" eingeführt, welche auf die Informationen der Ankertexte (siehe Tabelle 1) referiert, die

[4] Lokale Grammatiken kann man als „Landkarten der Sprache" bezeichnen [Mallchok, 2004], die einerseits Sequenzen von Wörtern, welche semantische Einheiten bilden, und andererseits syntaktische Strukturen beschreiben. Besonders auf dem Gebiet der lexikalischen Disambiguierung werden lokale Grammatiken verstärkt eingesetzt, welche üblicherweise durch einen endlichen Automaten bzw. einen Transduktor repräsentiert werden. In der Regel werden lokale Grammatiken in Form von Graphen [Paumier, 2004] visualisiert.

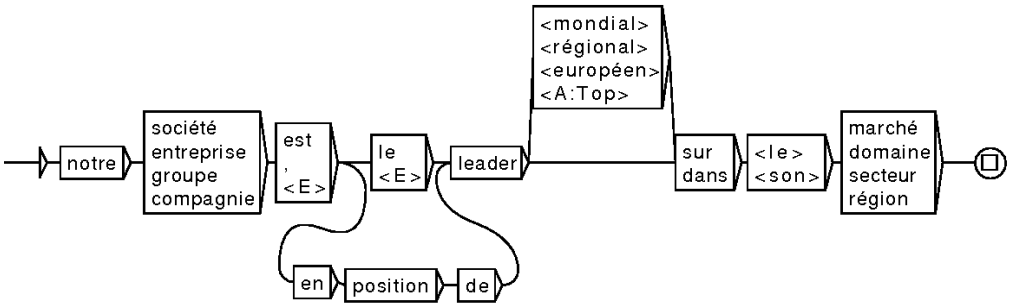


**Abbildung 2:** Lokale Grammatik, die den Kontext des Wortes *leader* modelliert

zu den offenen Stellenangeboten des jeweiligen Unternehmens führen, wie z.B.

- Wir stellen ein
- Unsere Stellenangebote

Bereits während der Lernphase konnten über 80 auf Stellenausschreibungen hinweisende Sequenzen dieser Kategorie zugeordnet werden.

### 2.2 Fokussierte Crawler

Eine weitere Vorgehensweise zum Auffinden von Stellenanzeigen konzentriert sich auf die typische Terminologie eines Stellenangebots, die einer Subsprache [Harris, 1968; 1988] gleichkommt. Dafür wurde noch während der Lernphase eine Reihe von Floskeln und komplexen Formulierungen gesammelt, die sich in zwei Typen einteilen lassen: Nominale Phrasen, welche die Stellenbeschreibungen semantisch strukturieren und sich in den Überschriften der einzelnen Abschnitte der Anzeige manifestieren, und Redewendungen, welche spezifische Verben oder Nomen der hier erwähnten Subsprachen beinhalten und nur im Kontext von Stellenangeboten Sinn ergeben.

Mit Hilfe der erwähnten Terminologie kann ein fokussierter Crawler nun entscheiden, ob eine Website in die Kategorie „Stellenanzeige“ einzuordnen ist, selbst wenn die HTML-Struktur des Dokuments keinerlei Aufschluss über eine mögliche semantische Gliederung der Anzeige gibt, oder statt allgemeiner Floskeln nur bestimmte Formulierungen darin auftreten.

Da die in Phase A (siehe Abbildung 1) gefundenen Stellenanzeigen uns nun als Volltext vorliegen, müssen sie anhand ihres Inhalts strukturiert werden, so dass ihr Informationsgehalt transparenter für den Jobsuchenden wird. Zu diesem Zweck werden einerseits die Daten des Stellenangebots über ihre semantischen Typen (z.B. Berufsbezeichnung, Firmenname) extrahiert, um später die Datenbank der Stellenanzeigen zu erstellen. Andererseits werden die Dokumente in Form von Vektoren dargestellt, welche die eben genannte Information enthalten, und auf denen letztendlich die Jobsuche effizienter gestaltet werden kann.

## 3 Informationsextraktion und Generieren von Dokumentrepräsentationsvektoren

Die zweite Phase des vorgestellten Systems (B) – der Hauptteil der vorliegenden Arbeit – behandelt die Informationsextraktion und die automatische Umwandlung der reinen Textform eines Stellenangebots in ein semantisch strukturiertes Dokument. Wir haben uns zu diesem Zweck auf die Erstellung einer bedeutenden Anzahl lokaler Grammatiken und elektronischer Lexika konzentriert, die es uns erlauben, die Datenbank der Stellenangebote automatisch zu füllen. Die Struktur der Datenbank kann als ein Formular betrachtet werden, über das die in jeder Stellenanzeige vorliegende Information strukturiert wird. Die gängigen Konzepte von Jobsuchmaschinen – auch neuerer Generation – erfordern hingegen ein manuelles Ausfüllen der entsprechenden Formulare.

| **Beispiel des auszufüllenden Formulars** | |
|---|---|
| Datum der Veröffentlichung | 22. Jan 2007 |
| Bewerbungsfrist | fin février<br>*(Ende Februar)* |
| Einstellungsdatum | mi-mars<br>*(Mitte März)* |
| Stellenbezeichnung | ingénieur d'étude en électromécanique<br>*(Elektromechanikprojektingenieur)* |
| Art des Vertrags | intérim à temps partiel :<br>1 à 2 jours/semaine<br>*(Zeitarbeit : 1 bis 2 Tage/Woche)* |
| Anstellungszeitraum | 8 mois renouvelables<br>*(8 Monate verlängerbar)* |
| Arbeitsort | sud-est de Paris<br>*(süd-westlich von Paris)* |
| Gehaltsvorschlag | selon profil<br>*(nach Profil)* |
| Referenz der Stelle | MOR34544/ing-21 |
| Arbeitserfahrung | expérience de 2 à 3 ans<br>dans un poste similaire<br>*(2 bis 3 Jahre Erfahrung in einem ähnlichen Beruf)* |
| gewünschte Ausbildung | de formation Bac+5<br>de type école d'ingénieur<br>*(Abschluss als Diplom-Ingenieur)* |
| Firmenname | CGF Sarl |
| Firmensitz | **Adresse :** 34 bis rue Berthauds, 93110 Rosny<br>**Tel :** 0 (+33) 1 12 34 45 67<br>**Fax :** 0 (+33) 1 12 34 45 68<br>**Email :** contact@cgf.fr<br>**Homepage :** http:///www.cgf.fr |
| Kontakt | Directeur des RH, Mr. Brice<br>*(Personaldirektor, Mr. Brice)* |
| Firmenbranche | Construction électromécanique<br>*(Elektromechanische Konstruktion)* |

**Tabelle 2:** Strukturiertes Stellenangebot in der Datenbank

Über ein im ersten Arbeitsschritt des Systems (siehe Abschnitt 2) gefundenes Dokument werden nun gewisse In-

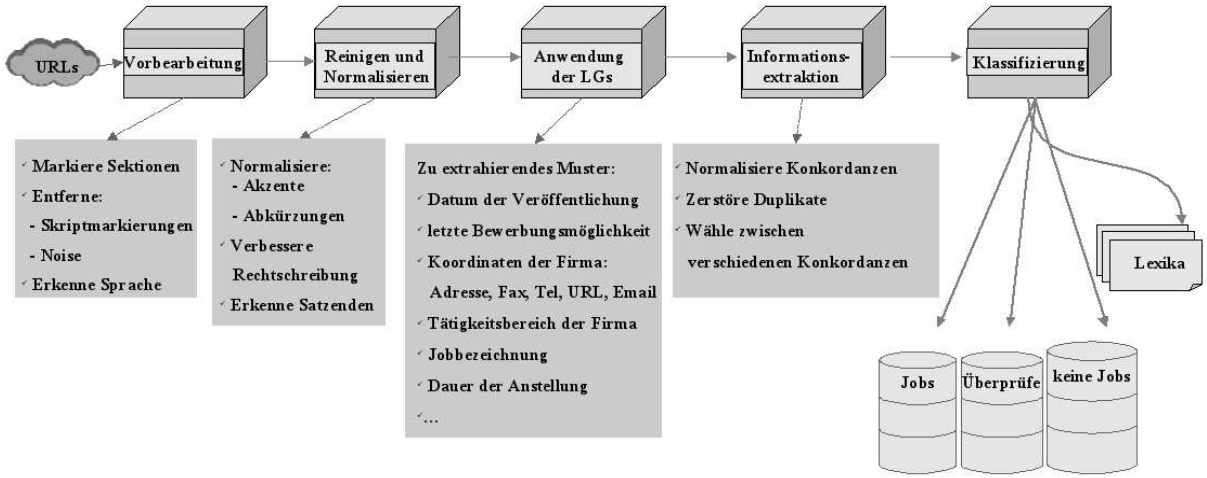


**Abbildung 3:** Vearbeitungsphasen eines potentiellen Stellenangebots

formationen automatisch aus dem Stellenangebot extrahiert, um damit ein Formular zu befüllen, wie es in Tabelle 2 abgebildet ist.

Für diese Transformation der ursprünglichen HTML-Dokumente in dieses Schema sind verschiedene Operationen nötig, deren chronologischer Ablauf in Abbildung 3 dargestellt wird.

In dieser Abbildung können fünf unterschiedliche Phasen unterschieden werden, auf die im folgenden näher eingegangen werden soll.

### 3.1 Vorverarbeitung

In diesem Schritt wird jedes Dokument mit semantisch-strukturellen Markierungen *(*[*TagMISSION*]*,* [*TagPROFIL*]*,* [*TagFORMATION*]*, usw.)* gelabelt. Bereits während der Lernphase wurden insgesamt 13 Klassen ausgearbeitet, die Floskeln oder Phrasen enthalten, welche die Rolle von Untertiteln in einer Anzeige spielen, und durch eines der eben genannten Tags repräsentiert werden. So spiegelt [*TagMISSION*] beispielsweise die Klasse der *Tätigkeiten* wieder und enthält u.a. folgende Floskeln:

- Ihr Aufgabenbereich
- Ihre Aufgabengebiete sind
- Das erwarten wir von Ihnen:

Im Verlauf der Systementwicklung konnten wir feststellen, dass Stellenangebote grundsätzlich auf drei verschiedene Arten geschrieben werden. Entweder werden Floskeln benutzt, um das Dokument in einem hohen Maß zu strukturieren, oder es handelt sich um reine textuelle Beschreibung, ohne deutliche Struktur, oder es sind sehr kompakte Anzeigen mit einem Minimum an Information.

Nachdem die semantische Struktur des zu untersuchenden Dokuments analysiert und markiert wurde, werden alle HTML-Formatierungen sowie Programmskripte (z.B. JavaScript, ActionScript) gelöscht.

Anschließend wird die Sprache des verbliebenen Textes durch einen Abgleich mit verschiedensprachigen Wörterbüchern bestimmt, wobei dieses Modul aber nur Aufschluss darüber gibt, wie groß die Wahrscheinlichkeit dafür ist, dass das getestete Fragment in dieser Sprache geschrieben ist. So können alle nicht als französische Stellenbeschreibungen erkannte Dokumente herausgefiltert werden, da sie für unsere Zwecke ohne Bedeutung sind.

### 3.2 Bereinigung und Normalisierung

Die Bereinigung der Daten besteht darin, die orthographischen Fehler sowie die fehlenden Akzente im Text zu erkennen und zu korrigieren, wobei nur die nicht ambigen Einträge berücksichtigt werden. Dies geschieht mit Hilfe einer Liste von häufigen Rechtschreibfehlern, die schon während der Lernphase auf einem großen Korpus von Stellenangeboten gesammelt wurden.

In der Normalisierungsphase wird versucht Abkürzungen zu identifizieren und diese durch ihre jeweiligen Originalformen zu ersetzen, z.B. *(ing.* $\mapsto$ *ingénieur, comm.* $\mapsto$ *commercial, usw.)*. Zu diesem Zweck wurde auch eine Liste von Abkürzungen extrahiert und ihren entsprechenden ausgeschriebenen Wortlauten zugeordnet.

Gegen Ende dieses Arbeitsschrittes verfügt man über einen bereinigten und normalisierten Text, auf dem nun die syntaktischen und lexikalischen Analysen durchgeführt werden können.

### 3.3 Anwendung der lokalen Grammatiken

Wir haben mehrere spezialisierte Lexika für einfache Wörter und für Mehrwortlexeme erstellt, welche die Terminologie der erwähnten Subsprache erfassen und auch konform mit dem DELA-Format [Courtois *et al.*, 1990; 1997] sind. Diese von uns eingehaltene Konvention erlaubt die Anwendung lokaler Grammatiken innerhalb der LGPL[5]-Software Unitex[6]. Bei dieser Plattform handelt es sich um ein Korpusverarbeitungssystem, welches es ermöglicht, mit elektronischen Lexika umzugehen und lokale Grammatiken in Form eines Finite-State-Graphen (Directed Acyclic Graph: vgl. Abb. 2) zu entwickeln und auf ein Korpus anzuwenden.

Um diese Graphen verarbeiten zu können, wird zunächst der Text in Unitex durch folgende Prozess-Pipeline geschleust:

1. *Convert*: Konvertiert den Text in Unicode (UTF-16LE).
2. *Normalize*: Normalisiert die Sonderzeichen, die Leerzeichen und die Zeilenumbrüche.
3. Satzendeerkennung

[5]GNU **L**esser **G**eneral **P**ublic **L**icense

[6]http://www-igm.univ-mlv.fr/ unitex/

4. Auflösung von Kontraktionen (z.B. *d'une* ↦ *de une*).
5. *Tokenize*: Tokenisiert den Text aufgrund des Alphabets der jeweiligen Sprache.
6. *Dico*: Führt eine lexikalische Analyse durch, indem die Lexika auf die Tokenliste des Textes angewendet, und jedem Wort seine möglichen grammatikalischen Kategorien zugeordnet werden.

Nach der Ausführung der lexikalischen Analyse folgt die Phase der Informationsextraktion mit Hilfe der lokalen Grammatiken. Für jeden Typ der zu extrahierenden Informationen (derzeit ca. 20 Stück[7]) wurden mehrere Grammatiken entwickelt, die iterativ bzw. kaskadiert [Friburger *et al.*, 2001] sowie mit unterschiedlicher Priorität ausgeführt werden. Beispielsweise wurden zur Extraktion der Berufsbezeichnungen die entsprechenden lokalen Grammatiken auf acht Prioritätsebenen verteilt. Die Graphen der Ebene $n + 1$ werden nur ausgeführt, wenn die Grammatiken der Ebene $n$ mit höherem Vorrecht keine Teffer liefern konnten.

### 3.4 Informationsextraktion

In der Phase, die wir als Informationsextraktion bezeichnen, geht es darum, die extrahierten Sequenzen zu normalisieren, sowie Duplikate und unvollständige Teilsequenzen zu entfernen. Nach einer Vielzahl von heuristischen Tests zeigte sich, dass es sinnvoll ist, jeweils den längsten Match zu bevorzugen. Obwohl man einräumen muss, dass in wenigen Fällen die falsche Entscheidung getroffen wird, blieben doch auf diese Weise die Verluste geringer.

Diese Entscheidungsphase ist notwendig, weil es oft vorkommt, dass Sequenzen mit verschiedenen Pfaden in den DAGs, die als Transduktoren fungieren, erkannt und dabei auf unterschiedliche Weise annotiert werden.

### 3.5 Klassifikation

Da es sehr selten vorkommt, dass ein Stellenangebot alle gesuchten Informationen gleichzeitig enthält, wurden Regeln ausgearbeitet, die in Abhängigkeit der jeweils erkannten Informationen im Dokument die Klassifikation in die Datenbank der Stellenangebote ermöglichen.

So wird die URL als Stellenanzeige klassifiziert und in der Datenbank durch die gefundenen Informationen indexiert, wenn einige der semantisch-strukturellen Floskeln, die Berufsbezeichnung und das Einstellungsdatum gefunden werden. Mit Hilfe einer benutzerfreundlichen Website kann jederzeit die annotierte Anzeige betrachtet werden. Dabei ist es auch möglich, die fehlenden Informationen manuell zu ergänzen. Denn die Benutzeroberfläche sollte eine schnelle Orientierung im Dokument ermöglichen und mit wenig Klicks können die fehlenden Felder des Formulars sowie die semantischen Wörterbücher um die farbig hervorgehobenen unbekannten Wörter erweitert werden.

## Anwendungen

An einem konkreten Beispiel sollen kurz typische Ergebnisse und die Qualität des Extraktionsprozesses anhand der entwickelten lokalen Grammatiken dargelegt werden.

In Abbildung 4 wird ein annotiertes Dokument nach Durchführung aller hier beschriebenen Verarbeitungsschritte gezeigt.

[7]Darunter fallen u.a. Informationen wie Berufsbezeichnung, Firmenname, Firmensitz, Arbeitsort und evtl. Gehaltsangaben.

Die erkannten, semantisch-strukturellen Floskeln, die in 13 Klassen[8] eingeteilt wurden, sind grün hervorgehoben. Im Beispiel konnten alle Einträge der fünf in der Stellenanzeige vorhandenen Klassen gefunden werden ( *„TAGCPN = Firmeninformationen“, „TAGPOSTE = Stellenbeschreibung“, „TAGEXP = Qualifikation des Kandidaten“, „TAGSALAIRE = Gehalt“, „TAGCONTACT = Firmenkontakt“*). Einige dieser Floskeln haben eine starke Filterfunktion, da sie ausschließlich in Stellenanzeigen vorkommen.

14 von den 20 gesuchten Informationen (rosa hervorgehoben) wurden mit mehr als 100 verschiedenen lokalen Grammatiken eindeutig identifiziert und mit den korrekten semantischen Tags assoziert, wie z.B.

`<PosteName>` = *(Berufsbezeichnung)*

`<Location>` = *(Standort)*

`<Duree>` = *(Dauer des Vertrags)*

`<Salaire>` = *(Angebotenes Gehalt)*

`<CPN>` = *(Firmenname)*

`<DomainOrg>` = *(Tätigkeitsbereich des Unternehmens)*

`<TypeContrat>` = *(Art des Vertrags)*

`<Reference>` = *(Stellenreferenz)*

In diesem Beispiel gibt es eine Sequenz, die zur gefundenen Konkordanz gehören sollte, aber nur teilweise und daher fehlerhaft durch die Grammatik extrahiert wurde. Der gefundene Arbeitsort wäre hier gemäß automatischer Extraktion *„de Paris“* obwohl es *„à l'extérieur de Paris“ (Großraum Paris)* sein sollte. In dieser primären Konfiguration wurde die vorhandene Semantik stark verändert und die Performanz unseres System verschlechterte sich deutlich: Wenn ein Jobsuchender nach einer Stelle in der Stadt Paris selbst suchen würde, bekäme er aufgrund des hier generierten Fehlers diese Stelle angeboten. Aber wenn der Arbeitsuchende auch außerhalb von Paris eine Stelle suchen wollte, hätte er nach *IDF oder Île de France* gesucht, was dem „Großraum Paris“ entspricht.

Gesuchte Sequenzen, die durch die lokale Grammatiken nicht gefunden wurden, sind gelb hervorgehoben. Im Beispiel handelt es sich um das „Einstellungsdatum“, das durch die zwei Ausdrücke *„vous devez impératiement être disponible sous 1 à 4 semaines“ (Sie sollten in 1 bis 4 Wochen verfügbar sein)* und *„. . . pourrait démarrer une mission très rapidement“ (. . . könnte die Tätigkeit sehr bald beginnen)* umschrieben wurde. Diese Ausdrücke stellen keine feste Datumsangaben dar, aber beinhalten die gewünschte Information. Diese beiden fehlenden Informationen sind inzwischen den entsprechenden lokalen Grammatiken hinzugefügt worden. In diesem Sinne werden alle lokalen Grammatiken ständig erweitert, was dank der verfügbaren Unitex-Oberfläche [Paumier, 2004] sehr schnell umgesetzt werden kann.

Ein weiterer Vorteil der lokalen Grammatiken ist die gute Übersichtlichkeit der schon beschriebenen und der noch fehlenden syntaktischen Strukturen. Falls eine Information nicht extrahiert werden konnte, kann ein Pfad aufgrund der hohen Modularität sehr schnell in den Graphen eingefügt werden.

[8]Mission, Tâches, Compétences, Qualités, Connaissances, Expérience et Formation, Durée, Date de début, Lieu, Description du poste, Salaire, Coordonnées du contact, Entreprise

URGENT ! <PosteName> DÉVELOPPEUR PERL - 94 - FREELANCE </PosteName>(H/F)
FR-IDF-ILE DE FRANCE

Descriptif :
Mon client, un éditeur de logiciel international, recherche de façon urgente un Développeur Perl.

[TAGCPN] La Société :
Mon client est un acteur majeur sur son marché travaillant avec les plus grands comptes internationaux. Suite à une surcharge importante, ils sont actuellement à la recherche d'un développeur Perl qui pourrait démarrer une mission très rapidement.
Mission située à l'extérieur <Location> de Paris </Location> (très facile d'accès par les transports en commun) pour laquelle vous devez impérativement être disponible sous 1 à 4 semaines.

[TAGPOSTE]Description de poste :
Vous devrez tout d'abord analyser plusieurs sites Web ainsi que leurs fichiers attachés puis vous aurez à charge de leur développement sous la derniere version de Perl.
Votre expertise technique, votre implication et votre motivation vous permettront d'évoluer au sein d'une équipe dynamique, pour un client qui apportera une forte valeur ajoutée à votre parcours.
Excellente opportunité de rejoindre une société très demandée, sur une mission de<Duree> 3 mois </Duree> avec de fortes possibilités de renouvellement.

[TAGEXP] Description des Candidats :
- Perl : 2 ans minimum
- Anglais est un plus
- XML : 1 an
- Html : 2 ans

[TAGSALAIRE] Tarif :
<Salaire> 290 à 330€/jour selon expérience <Salaire> .

[TAGCONTACT] Contact :
Si vous avez les compétences nécessaires, merci de me contacter très rapidement afin que je vous organise un entretien avec mon client.

<CPN> Computer Futures Solutions </CPN> est un acteur majeur sur le marché du recrutement et de la prestation de services au niveau Européen dans le domaine des <DomainOrg> technologies de l'information </DomainOrg> avec un chiffre d'affaires de plus de 220 Millions d'euros. Nous sommes présents dans les plus grandes capitales (Paris, Londres, Amsterdam, Bruxelles …).

Additional Information
Negotiable
Position Type:<TypeContrat> Full Time </TypeContrat>, Temporary / Contract / Project
<Reference> Ref Code: 391289 </Reference>

[TAGCONTACT] Contact Information
<Contact> <Prenom> Rudy </Prenom> <NomF> Nabet </NomF> </Contact>
<CPN> Computer Futures Solutions </CPN> - Paris
<Addresses> 33 RUE DE LA BOETIE, PARIS 75008 </Addresses>
Ph:<TEL> + 33 1 42 99 83 33 </TEL>
Fax:<FAX> + 33 1 42 99 83 00 </FAX>

**Abbildung 4:** Beispiel eines automatischen annotierten Stellenangebots

## 4 Evaluierung der Extraktionsergebnisse

Um die Qualität unseres Systems in der Erkennungsphase von stellentypischen Informationen zu demonstrieren, wurde ein kleines Testkorpus[9] bestehend aus ca. 1000 Stellenanzeigen manuell annotiert. Damit war es uns nun möglich, die Precision- und Recall-Werte für die automatisch gefundenen Resultate anzugeben und letztendlich auszuwerten, ob das System unsere Erwartungen erfüllt.

| Extrahierter Informationstyp | Precision | Recall |
|---|---|---|
| Berufsbezeichnung | 96.9 % | 93.3 % |
| Firmenname | 94.3 % | 90.6 % |
| Firmensitz (Adresse) | 93,0 % | 92,3 % |
| Gehaltsangabe | 97,1 % | 91,8 % |
| Arbeitsort | 98,0 % | 96,9 % |
| Im Durchschnitt | 95.86 % | 92.98 % |

**Tabelle 3:** Evaluationsergbnisse auf den Textkorpora

In Tabelle 3 wird deutlich, wie vielversprechend die Werte für Precision (durchschnittlich 95.86 %) und Recall (durchschnittlich 92.98 %) bei der Erkennung anzeigenspezifischer Entitäten sind. Für die hier vorgestellte Arbeit beschränkt sich die Auswertung auf fünf der wichtigsten von insgesamt 13 Informationsklassen. Daran wird ersichtlich, dass die eindeutige Identifikation des Firmennamens im Vergleich zu den anderen Kategorien noch die meisten Schwierigkeiten bereitet, aber dennoch um einiges qualitätiv besser ist als das, was marktführende Jobsuchmaschinen leisten.

## 5 Ausblick

Teile des hier konzipierten Systems einer optimierten Jobsuchmaschine können auch für andere Zwecke benutzt werden. Einerseits hält dieses System die Datenbank der Firmen und ihrer Websites immer auf dem aktuellen Stand und kann so für die verschiedensten Anwendungen von großem Nutzen sein. Andererseits konsultieren immer mehr Menschen das Internet, um beispielsweise einen Dienstleister oder Anbieter einer bestimmten Branche oder Region zu finden. Unser aktuelles Ziel ist es ein Klassifikationssystem zu entwickeln, das automatisch jedes Stellenangebot in die entsprechende Berufsbranche einordnen kann. Zu diesem Zweck soll eine Ontologie der Berufsbezeichnungen erstellt werden, welche es erlaubt, die Suche auf die semantischen Beziehungen zwischen den Anfragetermen zu erweitern.

## Literatur


[Bsiri, 2007] Sandra Bsiri. *Extraction d'information: Génération automatique d'une base de données d'offres d'emploi*. Doktorarbeit, LMU München, 2007.

[Courtois *et al.*, 1990] Blandine Courtois, Max Silberztein. Dictionnaires électroniques du français. In *Langues française 87*, 11-22. Larousse, Paris, 1990.

[Courtois *et al.*, 1997] Blandine Courtois et al. *Dictionnaire électronique des noms composés DELAC : les composants NA et NN* Rapport Technique du LADL 55, Paris, Université Paris 7, 1997.


[9] Siehe annotiertes Testkorpus unter: http://www.cis.uni-muenchen.de/ sandrab/DA/IE-Korpus1.html


[Ferber, 2003] Reginald Ferber. Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. dpunkt.verlag, 2003.

[Flury, 2005] Wolfgang Flury. *Information Extraction aus Online-Stellenanzeigen für eine Jobsuchmaschine*. Abschlussarbeit im Aufbaustudium „Computerlinguistik", LMU München, 2005.

[Focus RH, 2006] Focus RH. *Le guide des 500 meilleurs sites emploi*. Jeunes Editions, Levallois-Perret, 2006.

[Fondeur *et al.*, 2005] Yannick Fondeur, Carole Tuchszirer. Internet et les intermédiaires du marché du travail. In *La lettre de l'IRES, n° 67*. IRES - Institut de Recherches Economiques et Sociales, Noisy-le-Grand, 2005.

[Friburger *et al.*, 2001] Nathalie Friburger, Denis Maurel. Elaboration d'une cascade de transducteurs pour l'extraction des noms personnes dans les textes. In *TALN 2001*, Tours, 2-5 Juli 2001.

[Geierhos, 2006] Michaela Geierhos. Lokale Grammatiken. In *Grammatik der Menschenbezeichner in biographischen Kontexten*, 16-23, Magisterarbeit, LMU München, 2006.

[Glöggler, 2003] Michael Glöggler. Suchmaschinen im Internet. Springer, Berlin, 2003.

[Gross, 1993] Maurice Gross. *Local grammars and their representation by finite automata*. M. Hoey (Hrsg.): Data, Description, Discourse, Papers on the English Language in honour of John McH Sinclair, 26–38. HarperCollins, London, 1993.

[Gross, 1997] Maurice Gross. *The Construction of Local Grammars*. E. Roche und Y. Schabès (Hrsg.): Finite-State Language Processing (Language, Speech, and Communication), 329–354. MIT Press, Cambridge, Massachusetts, 1997.

[Gross, 1999] Maurice Gross. A bootstrap method for constructing local grammars. In *Contemporary Mathematics: Proceedings of the Symposium*, University of Belgrad, 229–250. Belgrad, 1999.

[Harris, 1968] Zellig S. Harris. Mathematical Structures of Language. In *Interscience Tracts in Pure and Applied Mathematics 21*, 230–238. Interscience Publishers John Wiley & Sons, New York, 1968.

[Harris, 1988] Zellig S. Harris. Language and Information In *Bampton Lectures in America 28*, 120–128. Columbia University Press, New York, 1988.

[Kowalski, 1997] Gerald Kowalski. Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers, Boston/Dordrecht/London, 1997.

[Mallchok, 2004] Friederike Mallchok. *Automatic Recognition of Organization Names in English Business News*. Doktorarbeit, LMU München, 2004.

[Paumier, 2004] Sébastien Paumier. *Manuel d'utilisation d'Unitex*, Université de Marne-la-Vallée, 2004.

[Senellart, 1998a] Jean Senellart. Locating noun phrases with finite state transducers. In *Proceedings of the 17th International Conference on Computational Linguistics*, 1212–1219. Montréal, Canada, 1998.

[Senellart, 1998b] Jean Senellart. Tools for locating noun phrases with finite state transducers. In *The computational treatment of nominals*. Proceedings of the Workshop, COLINGACL'98, 80–84. Montréal, Canada, 1998.