Schlussprüfung Morphologie und Lexikographie FS 09

Aufgabenstellung: Simon Clematide

Prüfung vom 2. Juni 2009
Institut für Computerlinguistik
Universität Zürich

Vorname _____ Matrikelnummer _____

Nachname _____

Für Studierende der folgenden Studiengänge:

☐ BA - Studiengang Computerlinguistik (Phil. Fakultät)

☐ BA - Studiengang Computerlinguistik und Sprachtechnologie (Phil. Fakultät)

☐ BA-Studierende (Wirtschaftswissenschaftliche Fakultät)

☐ Studierende des Nebenfachs Informatik mit Studienbeginn ab WS 04/05

☐ Multidisziplinfach (ETH)

☐ Andere:

Nur für Lizentiatsstudierende der Computerlinguistik als ein Fach aus der Phil. Fakultät:

Strasse: _____ Hauptfach: _____

PLZ/Ort: _____ E-Mail: _____

Aufgabe Nr.:	1	2	3	4	5	6	7	8	**Summe**
Punktzahl:	6	8	8	12	22	18	8	8	90
Davon erreicht:									

Note SU: _____ Note SP: _____

Endnote: _____ Bestanden: ☐ Ja ☐ Nein

Auf jedes Blatt mit Lösungen den Nachnamen schreiben! Viel Erfolg!

Wichtige Hinweise

Punkte-Maximum: 90 (pro Minute 1 Punkt)
Hinweis: Bitte schreiben Sie in einem überlegten und knappen, aber verbalen Stil (keine Stichwortsammlungen). Bei inhaltlichen Auswahlsendungen, wo einfach mal alles spontan hingeschrieben wird und Falsches wie Korrektes munter vermischt sind, behalte ich mir Abzüge vor. Das Zeitbudget ist so berechnet, dass man hälftig überlegen und schreiben kann.

6 **1. Wortbildung (6 Punkte)** Welche morphologischen Segmentierungen halten Sie für die Wörter „momentan“, „qualitativ“ und „verquält“ (wie in „Er markierte ein verquältes Lächeln“) für sinnvoll? Argumentieren Sie mittels analoger Beispiele.

Alle 3 Beispiele zeigen gewisse Probleme auf. Die folgenden Bemerkungen dazu sind weder als vollständig noch als alleinseeligmachend zu betrachten:

- momentan: Wortbildungmässig könnte es aus dem Substantiv „Moment“ und einem Adjektivsuffix „an“ gebildet sein. Allerdings finden sich kaum analog gebildete Adjektive, auch wenn es noch Adjektive gibt, welche auf „-an“ enden: „filigran“, „mundan“, „simultan“, „mediterran“. Die Herkunft aus dem lateinischen Adjektiv „momentaneus“, wie sie im Universalduden angegeben wird, macht etymologisch Sinn. Letztlich ist eine wortbildungsmässige Segmentierung im Deutschen aus synchroner Sicht nicht sinnvoll.
- qualitativ: Wortbildungsmässig könnte es analog wie „quantitativ“ aus den neoklassischen Formativen „qual“, „itat“ und „iv“ bestehen (GERTWOL). Oder aus „Qualität“ und „iv“ (Canoo). Für Nomenableitungen mit dem Adjektivsuffix „iv“ gibt es weitere Belege.
- verquält: Wortbildungsmässig kann es als departizipiales Adjektiv vom Verb „verquälen“ mit Verbpräfix „ver“ interpretiert werden, welches allerdings als „verquälen“ nicht existiert (vgl. Canoo fiktiver Eintrag). Eine Segmentierung von „ver“ macht Sinn, weil es ein häufig verwendetes Derivationsmorph bei Verben ist.

8 **2. Komposita-Bildung im Deutschen (8 Punkte)** Welche Probleme entstehen, wenn man für Erst- und Mittelglieder von Nominalkomposita einfach alle Stämme[1] (auch durch Derivation und Konversion entstandene) im Nominativ Plural und Singular sowie Genitiv Singular verwendet? Gewichten Sie die Probleme und geben Sie Beispiele!

Hier eine Auswahl von Problemen von Unter- und Übergenerierung:

- Ausserparadigmische Fugenelement können so nicht korrekt gebildet werden („Arbeitswut“). Schwerwiegend, da diese Klasse recht viele Wörter umfasst (-heit, -keit etc.)
- Übergenerierung: Es ist möglich, dass Formen akzeptiert werden, welche sehr ungebräuchlich sind. („Eitelkeitenwahn“). Schwerwiegend, falls Komposita generiert werden sollen.
- Nomen mit e-Schwund am Schluss können nicht vorhergesagt werden („Schulstunde“). Nicht so schwerwiegend, da es nicht so viele Wörter gibt, welche betroffen sind.

[1] Sehr unglückliche formuliert. Sollte eigentlich „Wortformen“ heissen.

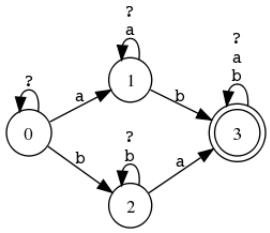
- Überanalyse: Wenn Abkürzungen und Einzel-Buchstaben (das „A“) zugelassen werden, werden viele ungewollte Analysen entstehen. Schwerwiegend.
- Konversionsresultate: Nominalisiere Infinitive und nominalisierte departizipiale Adjektive sollten ausgeschlossen bleiben. Bei Verben wird der Stamm zur Komposition verwendet („Singspiel“). Schwerwiegend.

8 **3. Infigierung (8 Punkte)** Ralf behauptet, dass es sich bei „zu“-Infinitiven wie bei „mitzukommen“ im Deutschen um einen Fall von Infigierung (d.h. Wurzel oder Stamm wird unterbrochen) handelt. Gemäss Lehmann kommt Infigierung in europäischen Sprachen aber sehr selten vor. Was meinen Sie? Begründen Sie Ihre Haltung.

Letztlich ist es eine Frage der Definitionen. Falls mit Stamm ein Flexionsstamm gemeint ist, wird er unterbrochen und Infigierung wäre durchaus zu sehen. Die Wurzel von „mitkommen“ ist „kommen“, welches vom „zu“ nicht unterbrochen wird.

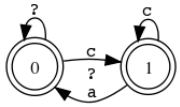
4. Reguläre Ausdrücke und Übergangsdiagramme (12 Punkte)

6 (a) Zeichnen Sie das Zustandsdiagramm eines minimalen Automaten, der die Sprache `[[$ a] & [$ b]]` akzeptiert. Überprüfen Sie Ihren Automaten, indem Sie sich Testbeispiele konstruieren!



Nicht zu vergessen ist, dass das `?` als Kantenbeschriftung die Zeichen im Sigma nicht umfasst.

6 (b) Schreiben Sie einen regulären Ausdruck, der folgenden Automaten ergibt:



Gesucht ist ein Automat, der kein a einlesen kann, dem nicht mindestens ein c direkt vorangeht. `[ a => c _ ]` oder `[\a* (c+ (a))]*` oder `[~ $a | c+ a]*`

5. Rechtschreibekorrektur à la GERSPELL (22 Punkte) Die Firma Lingsoft bietet eine Rechtschreibekorrektur an, welche folgendes Beispielverhalten zeigt:

4 (a) Gehen Sie davon aus, dass für GERSPELL ein Lexikon aller gültigen Wortformen des Deutschen zur Verfügung steht. Halten Sie die zulässigen Zeichenmodifikationen fest, auf Grund von denen für inkorrekte Zeichenketten die möglichen korrekten Wortformen vorgeschlagen werden.

Eingabe	Urteil	Vorschläge
Hauss	inkorrekt	Haus Haust Hause Hauses Hass Hausse Heuss Haugs Haubs
kleinr	inkorrekt	kleiner klein kleine
karz	inkorrekt	kurz karg
murz	unbekannt	
karzi	unbekannt	
bin	korrekt	

Um von einer inkorrekten auf eine korrekte Form zu kommen, darf entweder ein Zeichen eingefügt, gelöscht oder substituiert werden. Der 1. Buchstabe im Wort darf dabei nicht verändert werden („murz“ kann nicht zu „kurz“ korrigiert werden).

18 (b) Bauen Sie nun einen Transduktor namens `GENCORR`, der die entsprechenden **Vorschläge** auf Grund der bekannten Wortformen generiert. Gehen Sie davon aus, dass in der XFST-Variable `WF` ein Automat mit allen korrekten bekannten Wortformen vorliegt. Sie dürfen weiter davon ausgehen, dass Ihr Transduktor `GENCORR` nur dann verwendet wird, wenn eine Wortform nicht schon als korrekt erkannt wurde. D.h. es ist egal, was `GENCORR` für die korrekten bekannten Wortformen generiert.

```
xfst[1]: apply up murz
???
xfst[1]: apply up karz
kurz
karg
```

```
define CHAR [A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z|
             a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z];

define WF {klein}|{kleine}|{kleiner}|{kurz}|{karg}|{Haus}
|{Haust}|{Hause}|{Hauses}|{Hass}|{Hausse}|{Heuss}|{Haugs}|{Haubs};

# "_" wird als Marker für eine erfolgte Modifikation eingesetzt
define INS [ [..] (->) "_" CHAR || .#. ?+ _ ]; # Einfügen
define DEL [ CHAR (->) "_"      || .#. ?+ _ ]; # Löschen
define SUB [ CHAR (->) "_" CHAR || .#. ?+ _ ]; # Substituieren

define GENCORR WF
   .o. [INS|DEL|SUB]  # Modifikationen
   .o. [$. "_"]       # Muss genau 1 Marker enthalten
   .o. ["_" -> 0];    # Marker löschen

read regex GENCORR;
```

18 **6. Diminutivbildung (18 Punkte)** Geben Sie eine regelbasierte Implementation in xfst, welche Ableitungen wie „Mann“ „Männchen“, „Frau“ „Frauchen“, „Hund“ „Hundchen“/„Hündchen“, „Katze“ „Kätzchen“, ’„Maul“ „Mäulchen“, „Rubin“ „Rubinchen“ aus Nominalstämme machen kann. Überlegen Sie sich eine Lösung, wo nur für die Ausnahmefälle eine explizite Information in Stämme eingefügt werden muss und konzise Repräsentationen verwendet werden. Halten Sie fest, was Ihre Lösung nicht kann, wenn Sie nicht alle Fälle erschlagen können.

Den Beispielen entsprechend scheint es notwendige, optionale und nicht-mögliche Umlautung zu geben im Deutschen beim Diminuitivsuffix „chen“. Wenn der notwendige Fall als Default genommen wird, müssen die beiden andern Fälle markiert werden. Wenn diese Stämme nicht in separate Lexika gesteckt werden, müssen Marker im Stamm die Anwendung der Regel steuern (hier: `-UD` für kein Umlaut im Diminutiv). Im Folgenden ist eine

```
define VOC a|u|o|i;
define CON b|c|d|f|g|h|j|k|l|m|n|o|q|r|s|t|v|w|x|y|z;

define CHEN [[..] -> "+DIM" {chen} || _ .#. ];
define UML [a @-> ä,{au}@->{äu},u@->{ü} || _ [~$[VOC|"-UD"]] "+DIM"];
define ETILG [e -> 0 || _"+DIM"]; # E-Tilgung ist verbesserungsfähig!
define CLEAN ["+DIM"|"-UD" -> 0];

define DIMINUTIVA  LEXICON .o. CHEN .o. UML .o. ETILG .o. CLEAN;
```

8 **7. Vorteile und Nachteile von Finite-State Ansätzen (8 Punkte)** Sie haben selbst Erfahrungen damit gesammelt. Diskutieren Sie Vorteile und Nachteile dieses Ansatzes aus Ihrer Sicht.

Aus meiner Sicht sind die Vorteile: Generierung/Analyse im gleichen System ohne Effizienzverlust; Effizienz der Verarbeitung in Speicheranforderung und Rechenzeit; Generelles Framework für unterschiedliche Bedürfnisse (Konkatenation, Ersetzung, Komposition); Saubere und einfache theoretische Grundlage.

Nachteile: Verschiedene Frameworks mit unterschiedlichen Syntaxen und z.T. gewöhnungsbedürftiger Syntax; gewisse Konstrukte sind trotz generell effizienter Verarbeitung doch noch zu anforderungsreich; die Konstrukte der „normalen“ Programmiersprachen fehlen manchmal;

8 **8. Termextraktion (8 Punkte)** Beschreiben Sie je die Idee hinter einem linguistischen und quantitativen Ansatz der Termextraktion.

Siehe Skript.