Sommersemester 2003 • Institut für Germanistik I

Vorlesung
Computerphilologie

Themenfeld
Lexikalische Repräsentation

„Wie kann man Wörter und Beziehungen
zwischen ihnen beschreiben und erheben?“

Themenkreis Lexikalische Repräsentationen

Denkbare Themen:

- Morphologische Verarbeitung
- Lexikondefinition
- Lexikalische und semantische Sammlungen
- Bearbeitung von Wörtern in einem Korpus
- Stylometrie
- Lexikalische Autorenzuordnung
- Semantische Ähnlichkeitsberechnungen
- Konkordanzen und Thesauri

Morphologie

Zentrale Aufgaben:

- Zerlegung von Wörtern
- Lemmatisierung und
- Funktionserkennung

Was ist Morphologie

(zur Wiederholung)

Definitionen:

Morphologie ist die Lehre von den Klassen und Formen der Wörter einer Sprache;

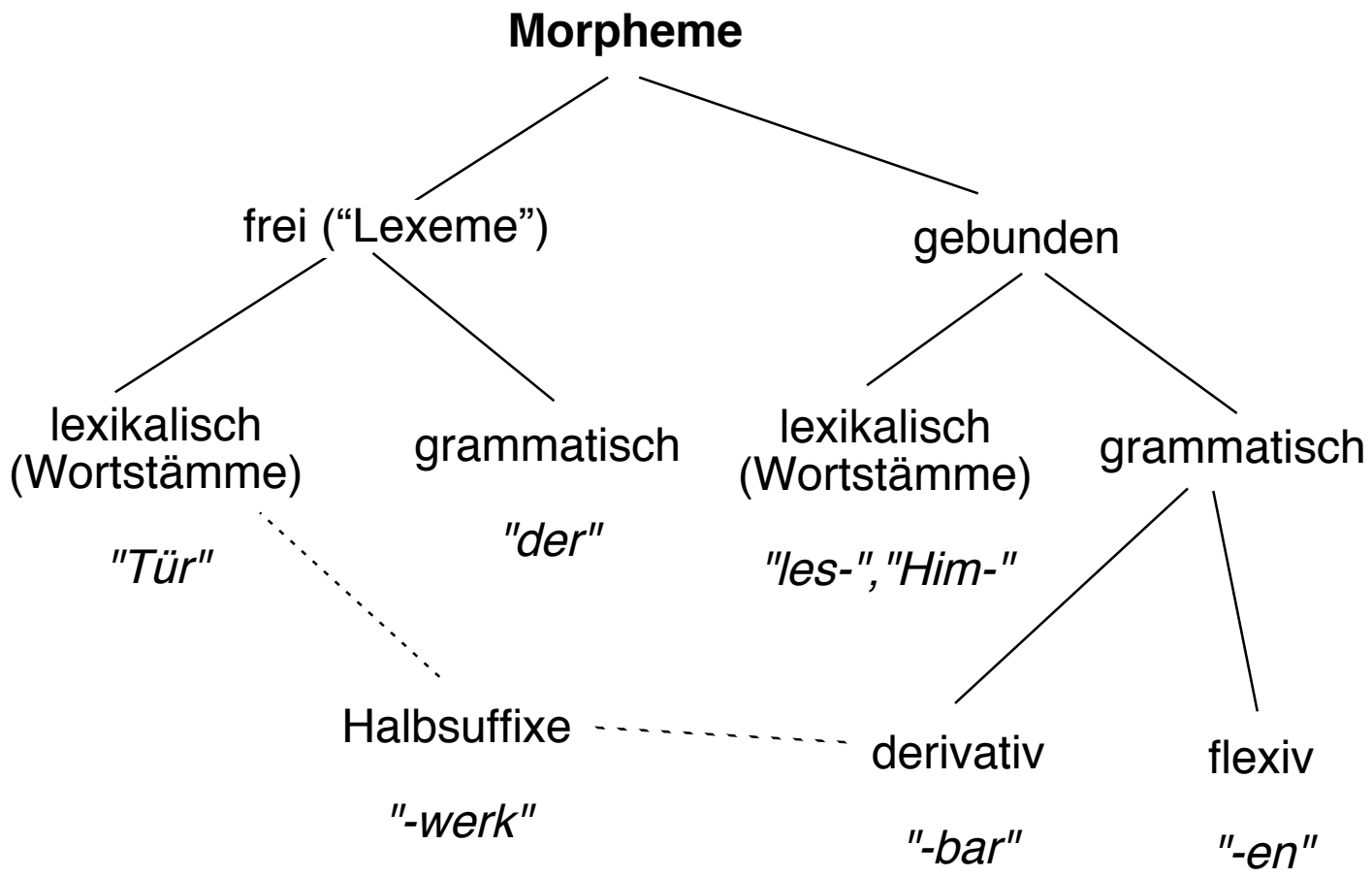
Spezieller: Lehre von den Wortbestandteilen, den Morphemen.

Diese Beschreibungebene der Sprache befaßt sich mit:

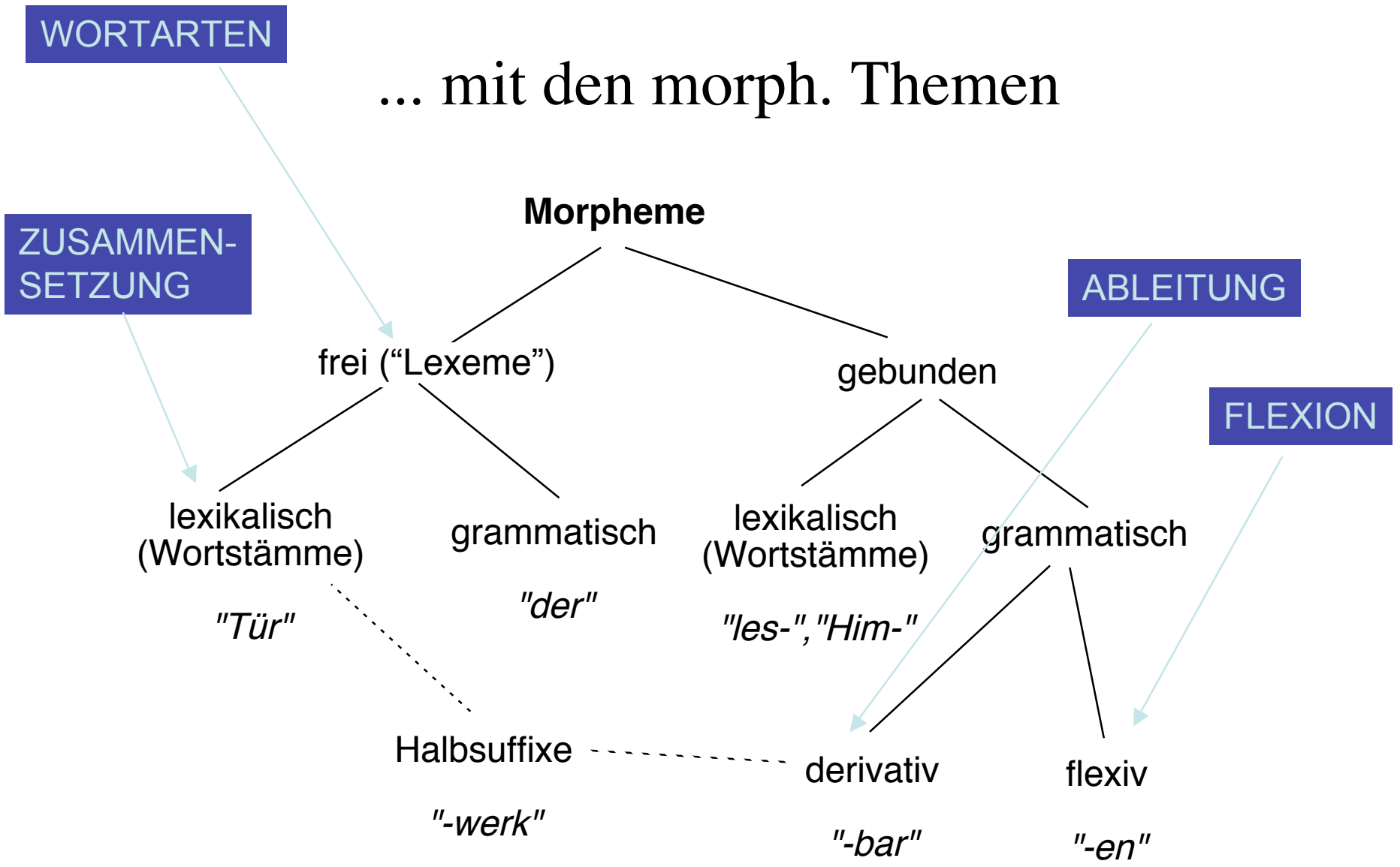
- Wortklassen
- Flexion
- Ableitung
- Zusammensetzungen (in Grammatiken bisweilen auch als "Wortbildung“ getrennt von der Morphologie behandelt)

jeweils unter syntaktischem und semantischem Aspekt

Morphemklassifikation



... mit den morph. Themen



Wozu morphologische Verarbeitung?

- Entlastung des Lexikons (Vollformenlexikon) von Flexionsformen und Zusammensetzungen ⇒ besonders wichtig für das Deutsche
- Entlastung des Lexikons (Stammlexikon) von phonologisch/graphemischen Phänomenen)
- Entlastung weiterer Verarbeitung auf allen Ebenen durch Anwendung nur auf lemmatisierte Formen
- Suffix-Information (Flexion) ist für die Syntax wichtig:
  - bei Nomen: Kasus, Genus, Numerus,
  - bei Verben: Person, Numerus, Modus, Tempus, Genus verbi (Aktiv/ Passiv)
  - bei Adjektiven: Kasus, Genus, Numerus, Steigerungsstufen
- Analyse spontaner Zusammensetzungen und Ableitungen,
- Trennbare und nichttrennbare Partikeln (im Deutschen),
- Generierung von Zusammensetzungen (und Ableitungen?),
- Erkennung von Wortklassen und anderen Funktionswechseln

Hauptprobleme der Verarbeitung

- Formal:
  - Vollformenlexikon vs Analyseprozesse
  - Anbindung an Lexikon oder Syntax i.e.S. oder eigene Morphologie
  - Keine Trennung üblich (wie bei Syntax) in morphologischen Algorithmus und Morphologie (Grammatik)
- Semantisch:
  - Entdeckung von semantischen Relationen bei Zusammensetzungen
    - "Jägerschnitzel" <-> "Schweineschnitzel"
  - Generierung von Zusammensetzungen und Zusammenbildungen
    - "Kosten der Reise" --> "Reisekosten"
  - Lexikalisierung vs Produktivität
    - (Entdeckung von Neubildungen bei produktiven Bildungsmustern)
  - Metaphorik
    - "Buchstabenkiller"

Morphologie und Linguistik

Relevanz für inhaltliche Verfahren:

• Wortklassen	hoch, weil oft nur syntaktisch erkennbar
• Zusammensetzung	hoch, weil die Analyse nur semantisch oder aus Kontext möglich ist
• Ableitung	hoch, weil reguläre Suffixe, aber hoch, weil die Analyse nur semantisch oder aus dem Kontext möglich ist
• Flexion	gering, weil durch Endungsbaum beschreibbar

Linguistische Quellen

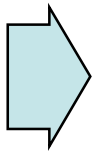
- Lexikon
  - Gesamtlexikon
  - Teillexika für unselbständige Lexeme
  - Morphotaktische Regeln

Anfangsüberlegung: Einfacher Algorithmus für Wortsegmentierung

Morphemische Methode

- Ab Wortanfang kürzeste Übereinstimmung mit einem Lexikoneintrag suchen. Den Rest gegen Lexikon prüfen. Aber das Verfahren ist
  - zeitaufwändig und umständlich, weil Länge unbekannt
  - unbekannte Morpheme in Zweitstellung führen zum Fehler
- Gegenbeispiele:
  - Rotzunge → Rotz Fehler, weil unge kein Lexem oder Suffix
  - Rotzunge → Rot Erfolg, weil Zunge Lexem

  Aber:
  - Spargelder → Spargel kein Fehler, weil der ein Lexem
  - Staubecken → ist ambig



- Schwierigkeit: Keine morphotaktische Information!

Wortsegmentierer durch N-Gramm-Tabelle

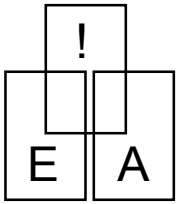
	A	B	C	D	E
A	A	A	AE	AE	!
B	A	A	!	AE	A
C	A	!	!	!	A
D	AE	!	!	E	AE

B

Digramme im Wort ab Pos 4 schrittweise gegen Digrammatrix klassifizieren:

- E = mögliches Wortende (z.B. -lz-)
- A = möglicher Wortanfang (z.B. -ro-)
- ! = im Wort unzulässig (z.B. -zr-)

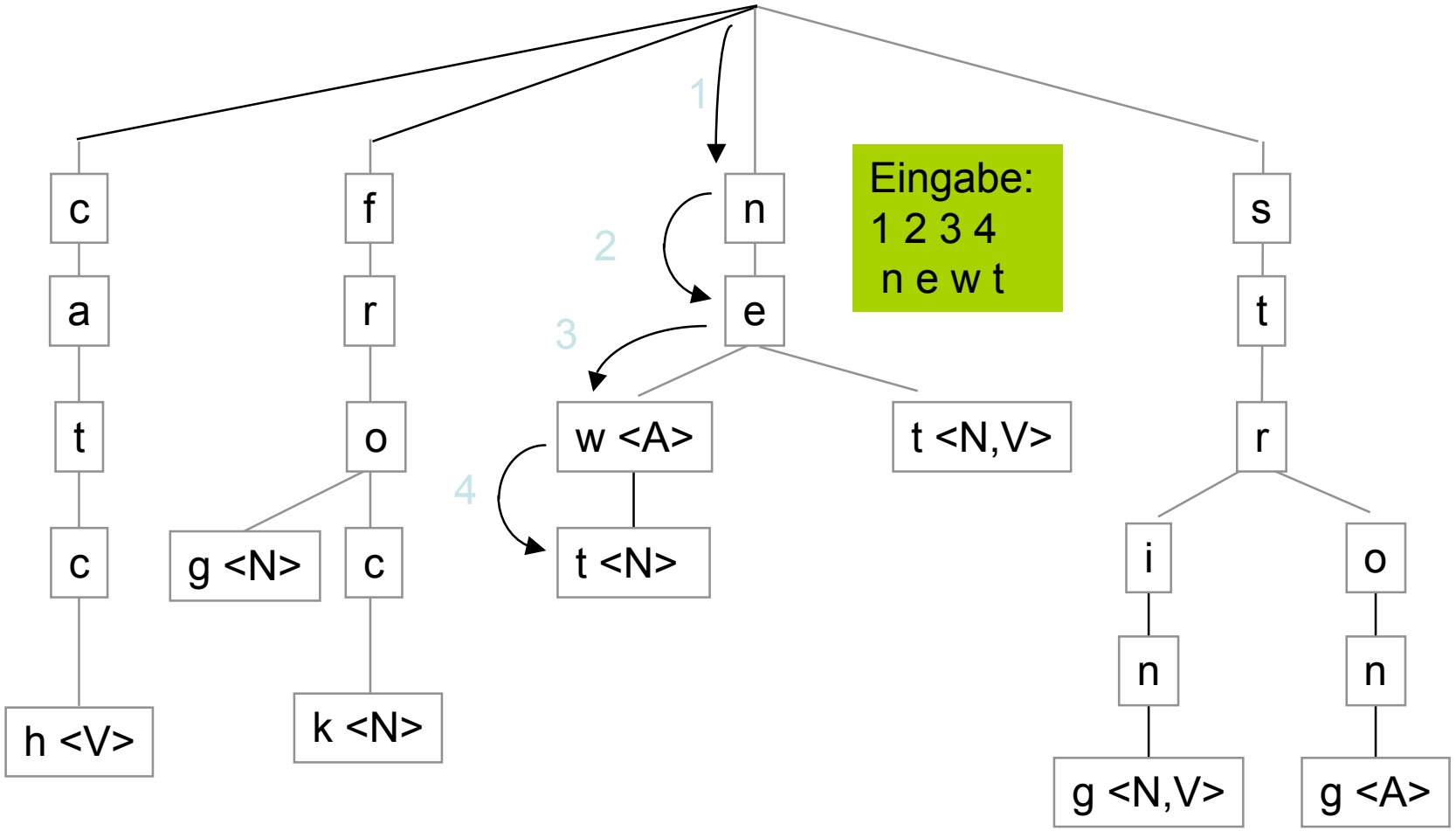
- Im Fall ! liegt eine Wortfuge im Digramm oder es ist ein unbekanntes Wort.
- Die Digramme an der Position - 1 müssen nach E klassifiziert sein, an der Position +1 nach Klasse A.



- `Endbetrag`

Erreichbar sind damit ca 95% korrekte Trennungen. Nur die Konsistenz der Reste ist ggf. gegen das Lexikon zu überprüfen.

Buchstabenbaum (Entscheidungsbaum)



Morphologiesystem MORPHY

http://www-psycho.upb.de/zinki/Kognition.html

- Für deutsche Sprache:
  - morphologische Analyse
  - statistischer PoS-Tagger
  - context-sensitiver Lemmatizer
- Das System kann auch zum Deutschlernen benutzt werden.
- Platform: Windows95/NT oder höher
- Nicht-ASCII-Zeichen werden nicht unterstützt

Morphologiesystem MORPHY
Lexikon

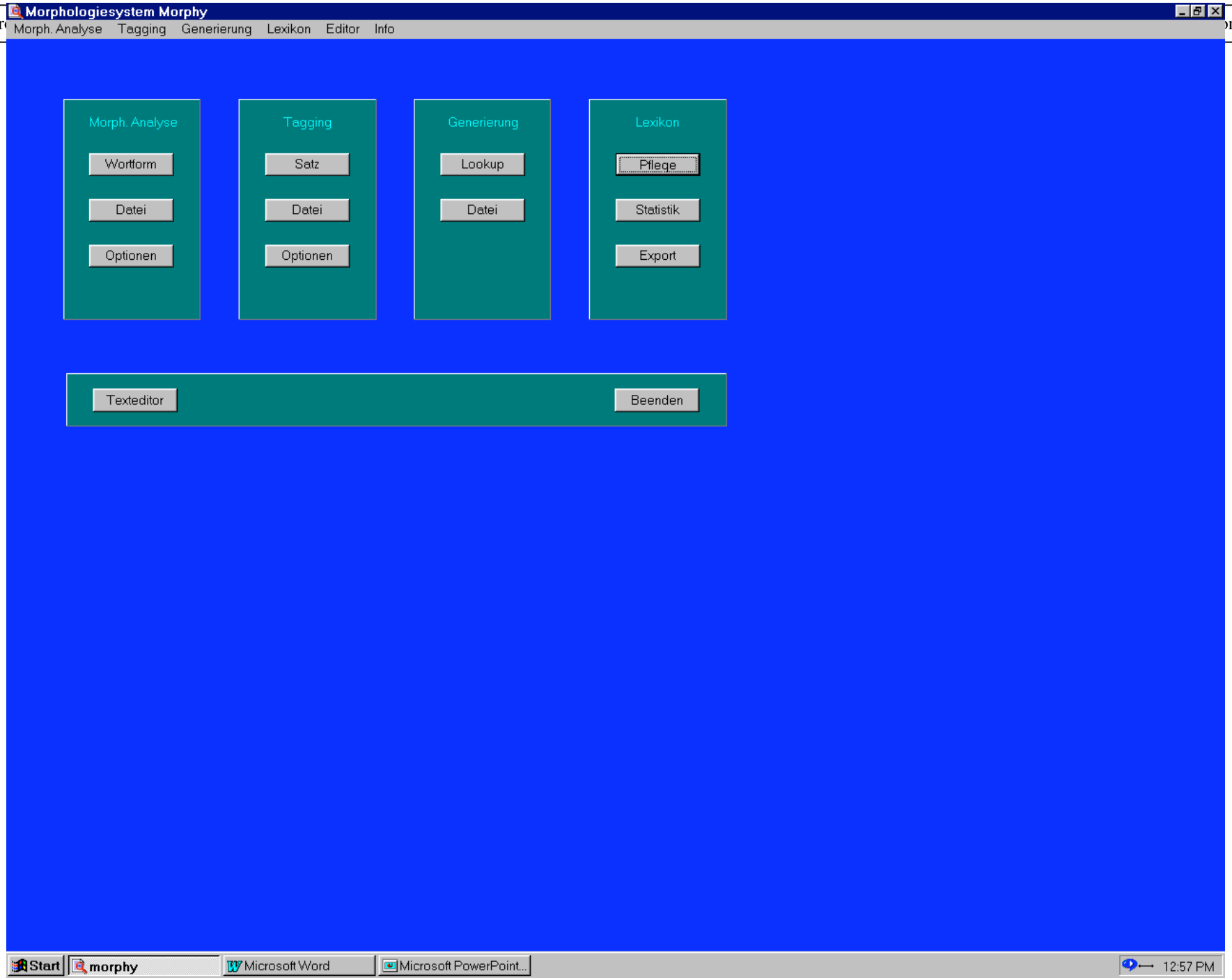
- Stammlexikon
- Das Lexikon ist in kleinere Lexika unterteilt, die jeweils eine Wortklasse umfassen.
- Jede Wortklasse besitzt eine eigene Datenstruktur, die alle Informationen zur Generierung enthält.
- Für hochfrequente Wörter ist zusätzlich ein kleines Vollformenlexikon eingerichtet.

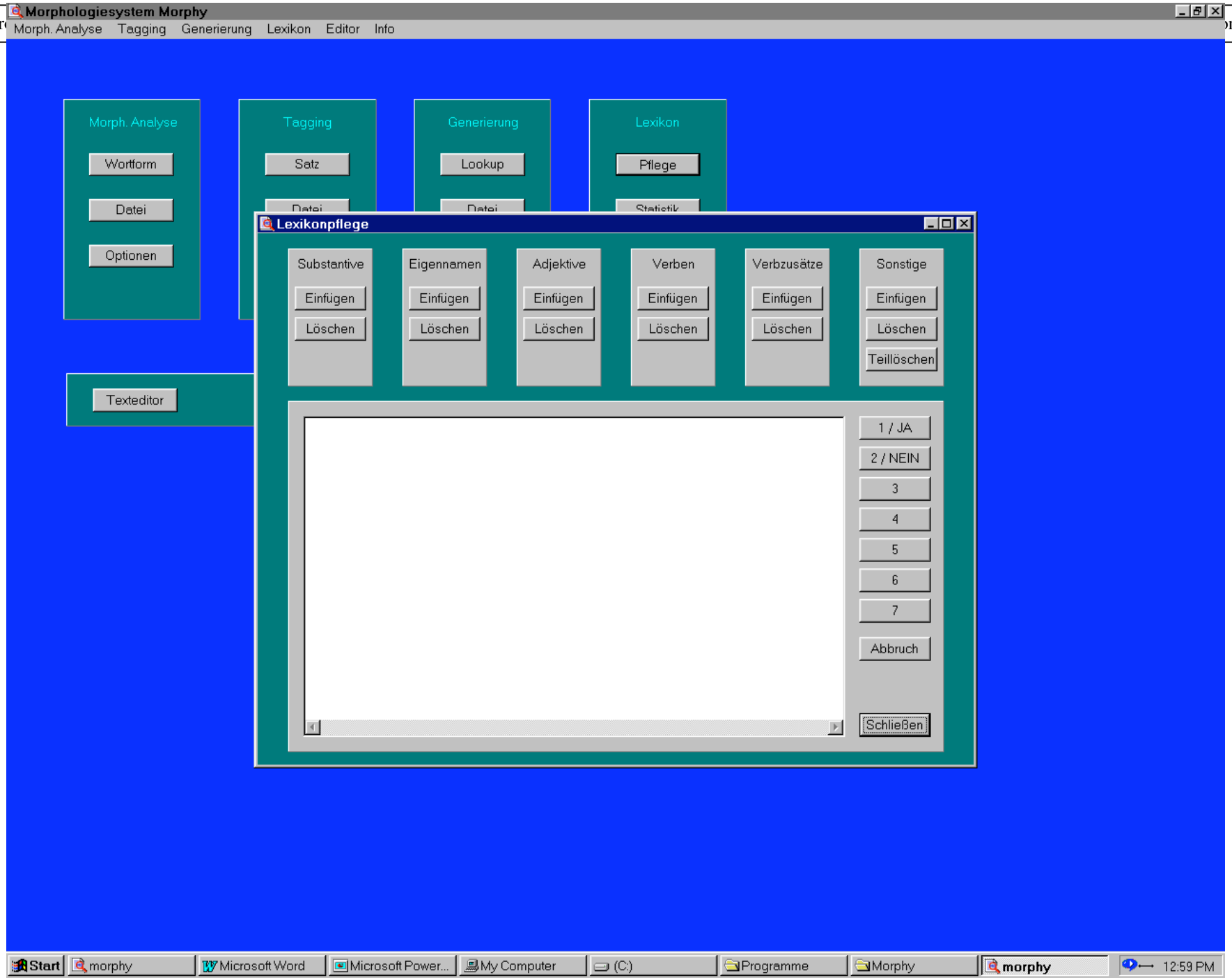
Morphologiesystem MORPHY
Wortklassen (I)

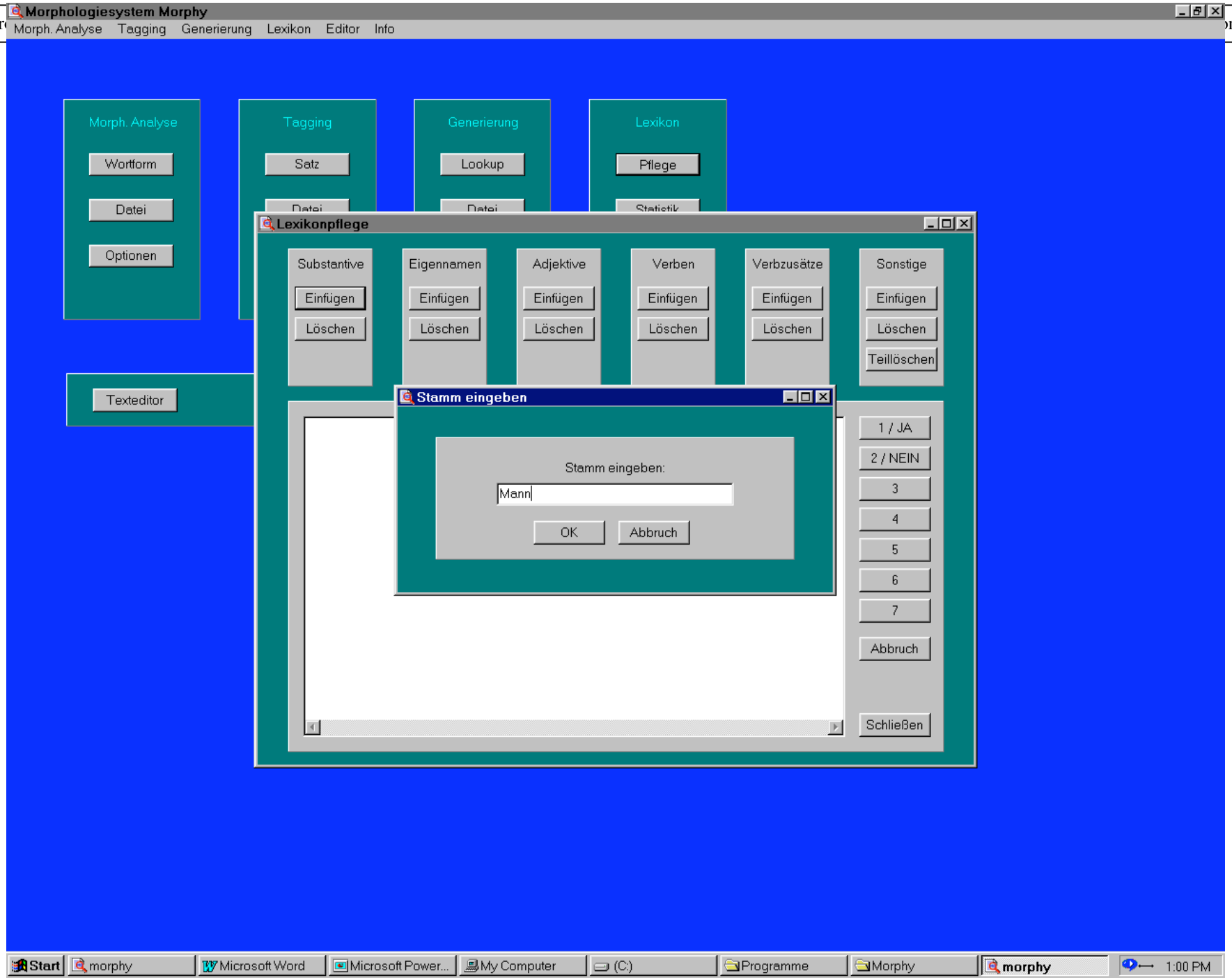
- Substantive:
  - 62 mögliche Deklinationsklassen (in einer Tabelle gespeichert)
  - Die folgende Information wird zu einem Stamm gespeichert:
    - Deklinationklassen (enthält Genus),
    - Pluraländerung:
      - “ß” durch “ss”
      - umlautender Vokal
- Adjektive:
  - Deklinationsklasse (legt das Deklinationmuster fest)
  - best./unbest./ ohne Artikel
  - Komparativ / Superlativ

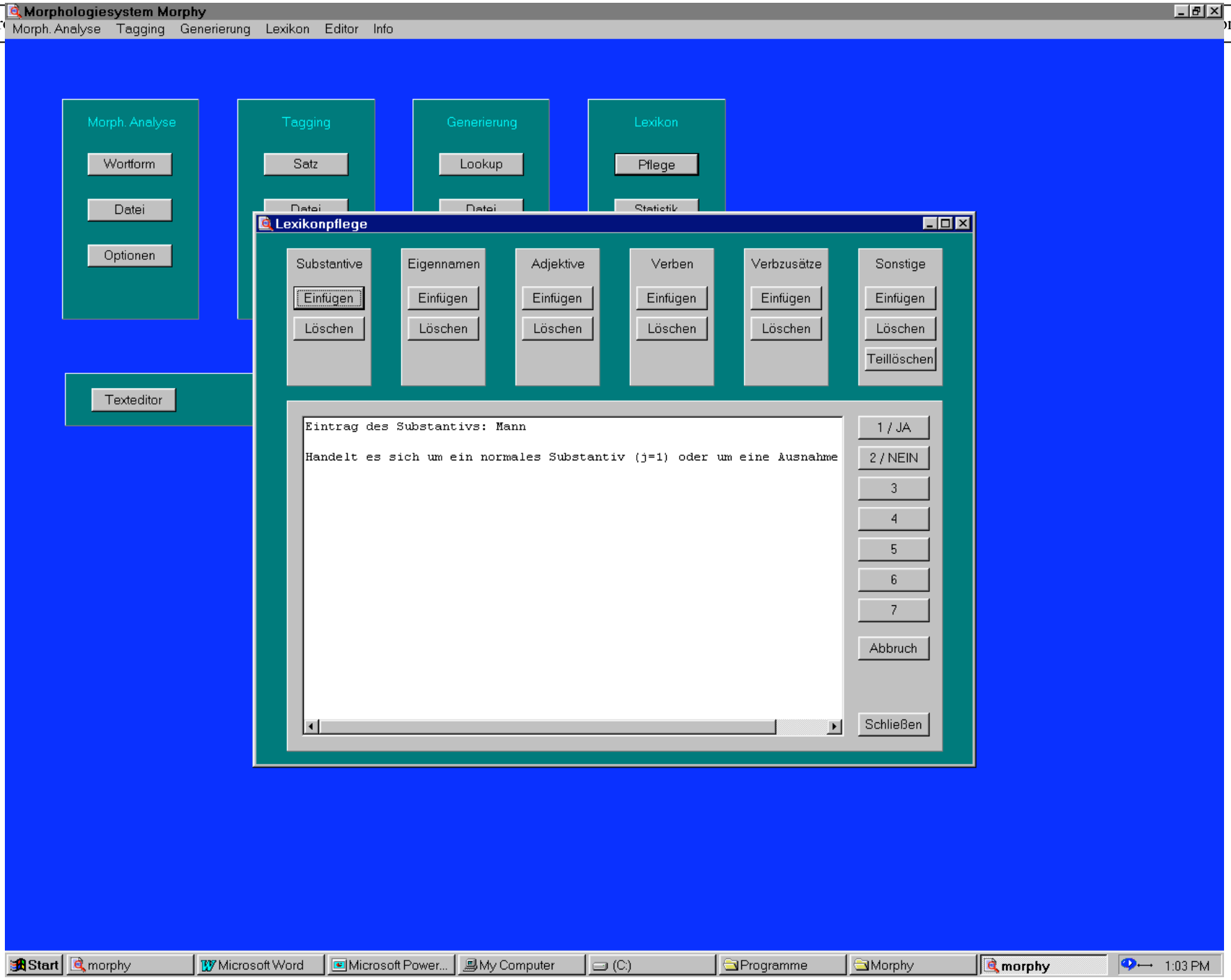
Morphologiesystem MORPHY
Wortklassen (II)

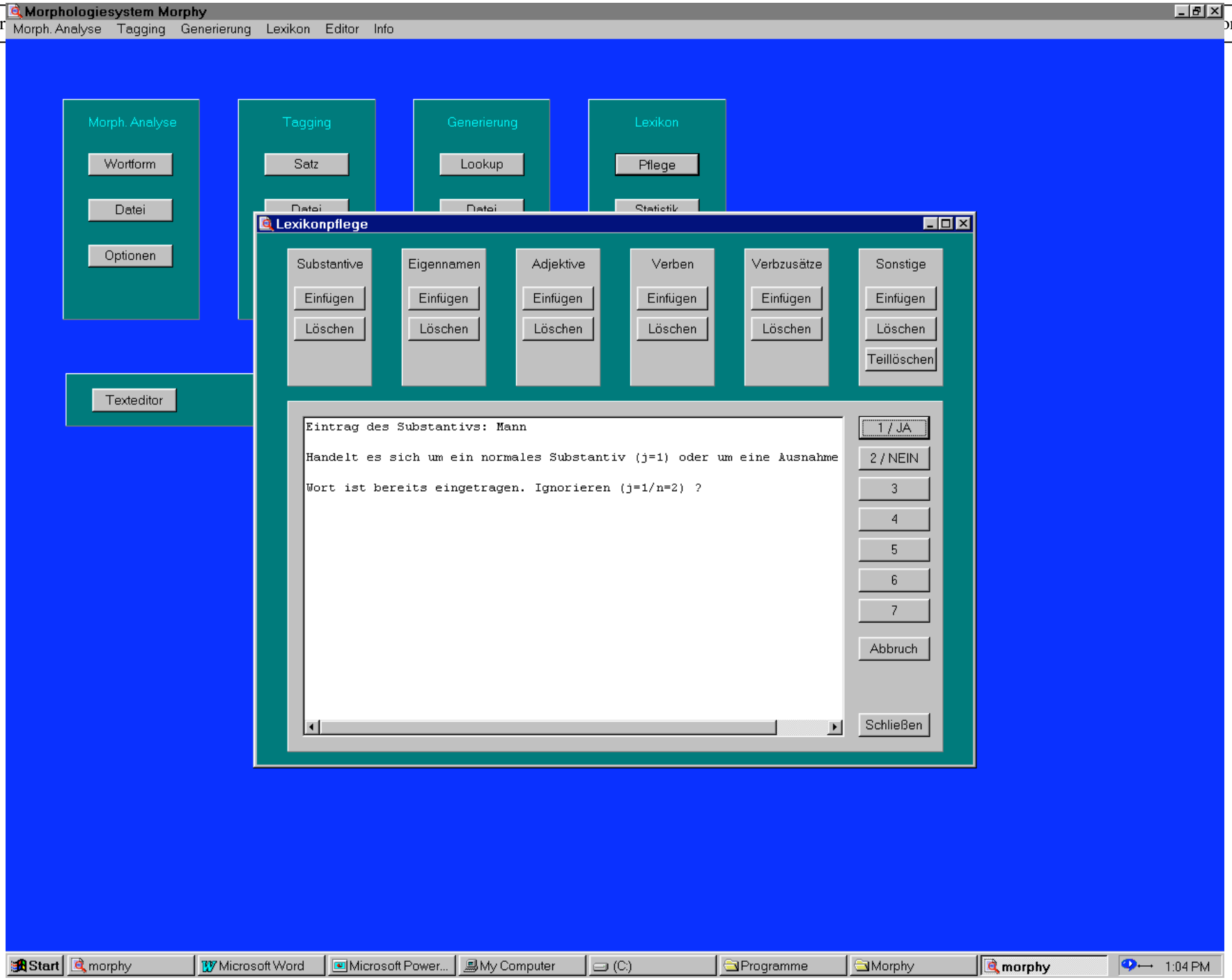
- Schwache Verben:
  - Konjugationsklasse
- Starke und gemischte Verben
  - die 7 markanten Formen
- Eigennamen
  - Genitiv
  - Genus
  - mit/ohne Artikel
- Für die übrigen Wortklassen (Partikel, Interjektionen) wird die Morphologie von Hand eingegeben

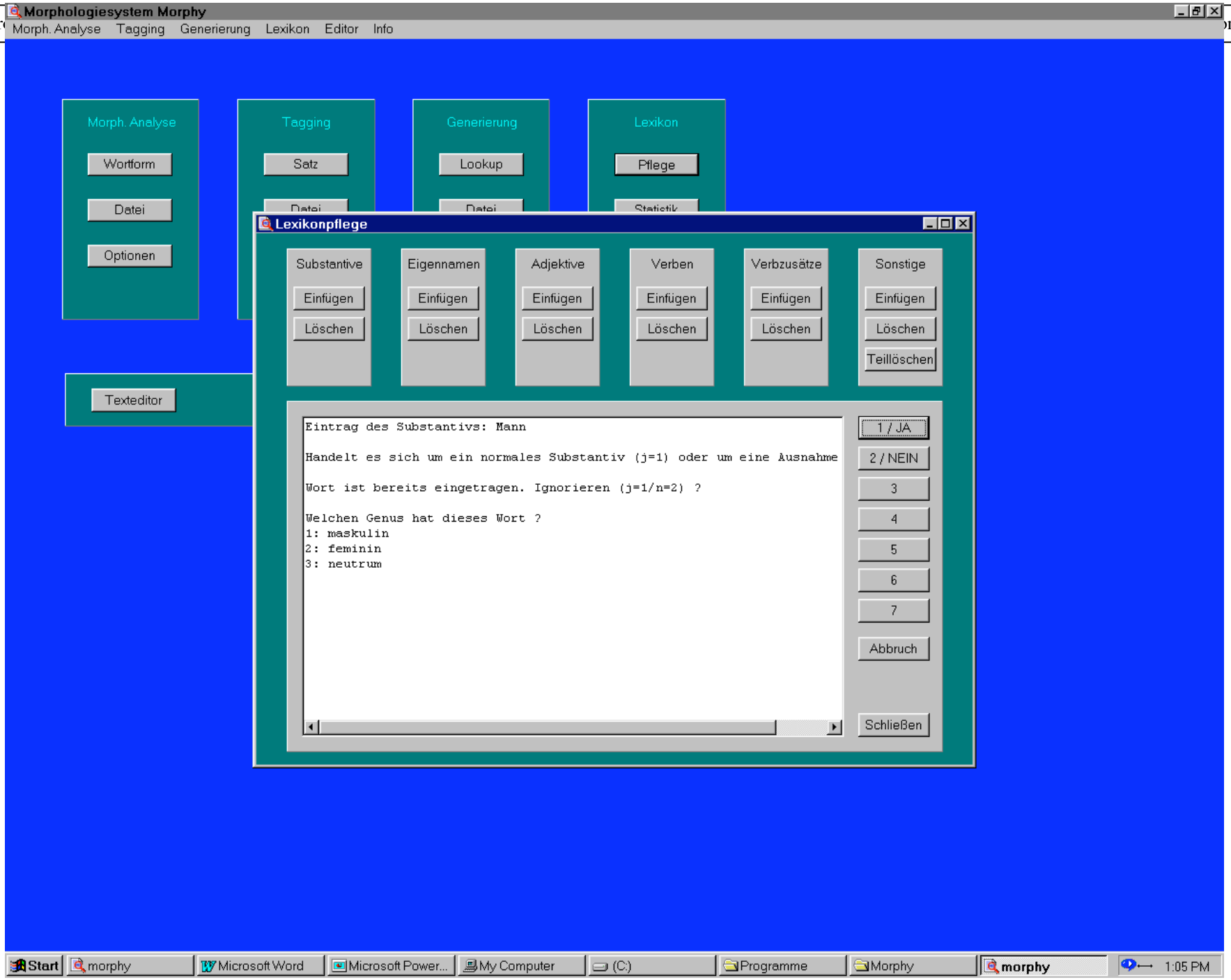

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Optionen
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Export
Texteditor
Beenden
Start
morphy
Microsoft Word
Microsoft PowerPoint...
12:57 PM

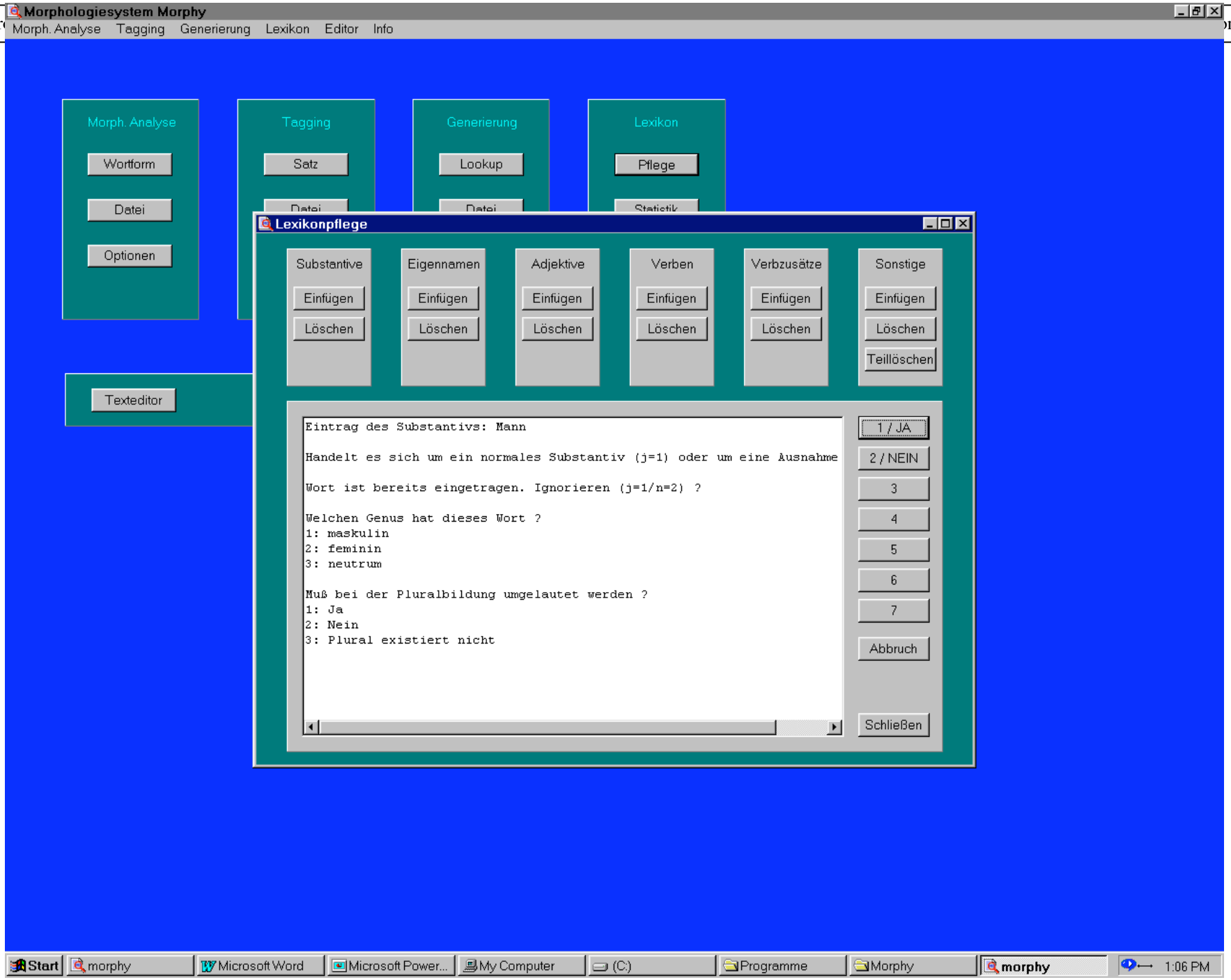

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
12:59 PM

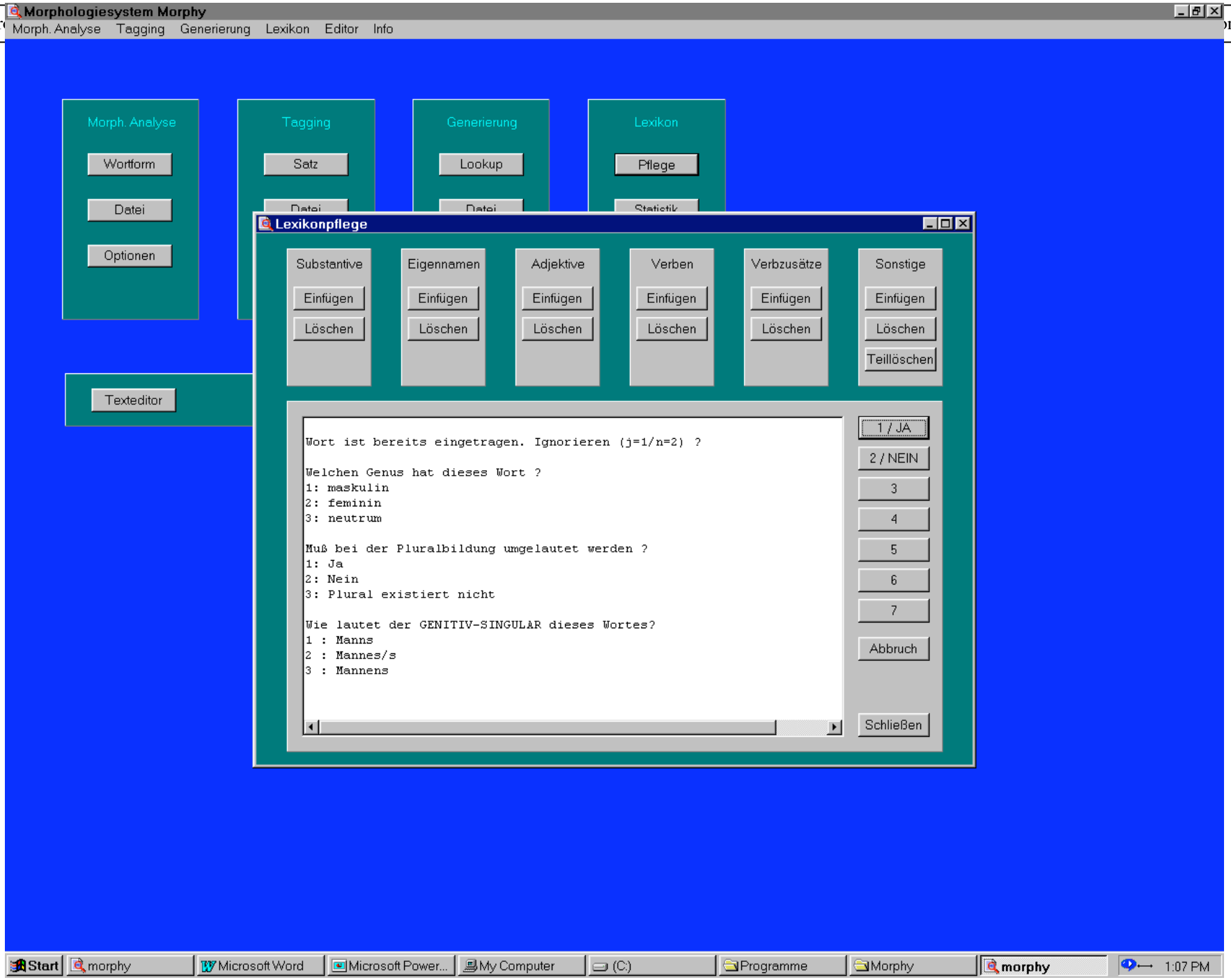

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
Stamm eingeben
Stamm eingeben:
Mann
OK
Abbruch
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:00 PM

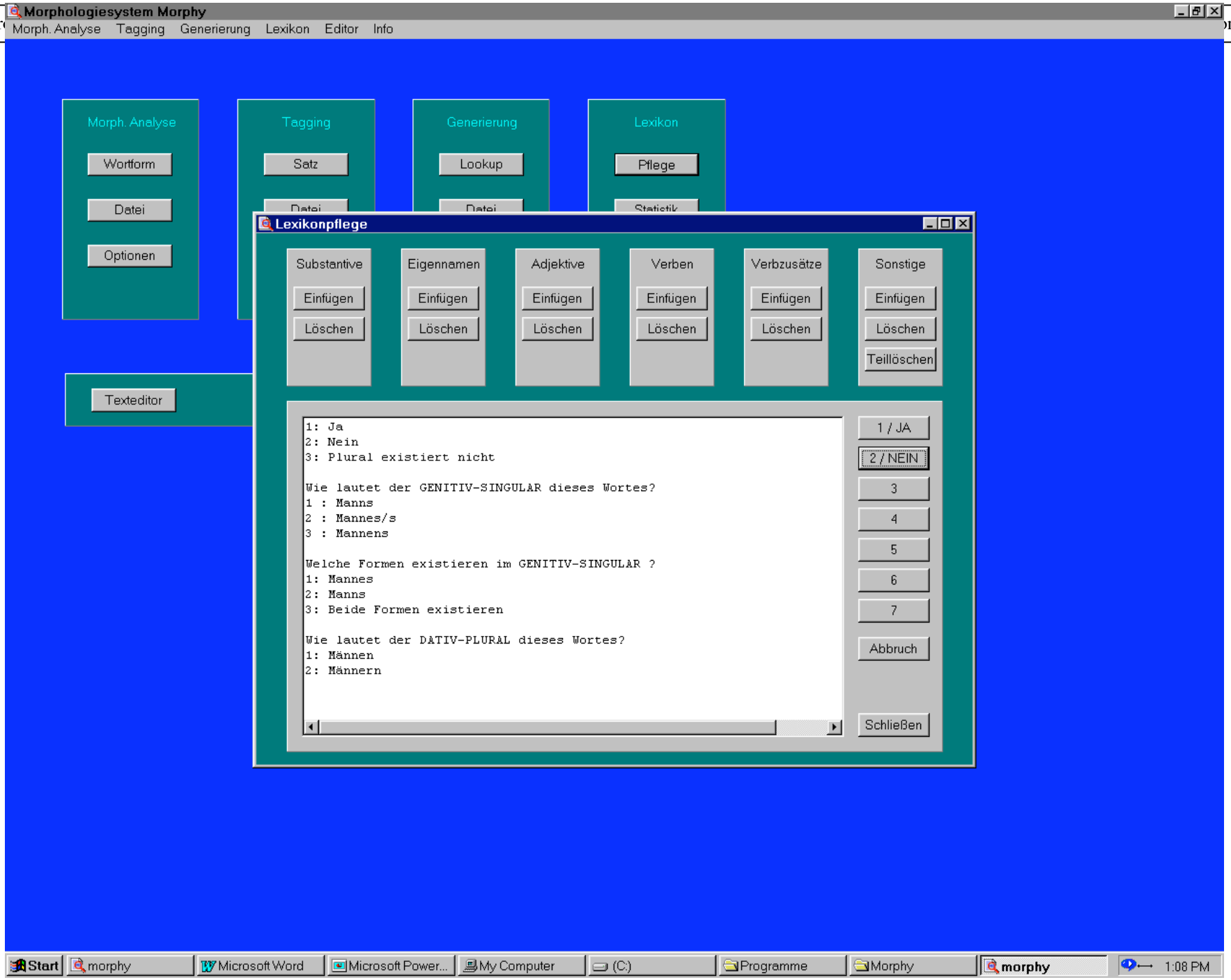

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
Eintrag des Substantivs: Mann
Handelt es sich um ein normales Substantiv (j=1) oder um eine Ausnahme
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:03 PM

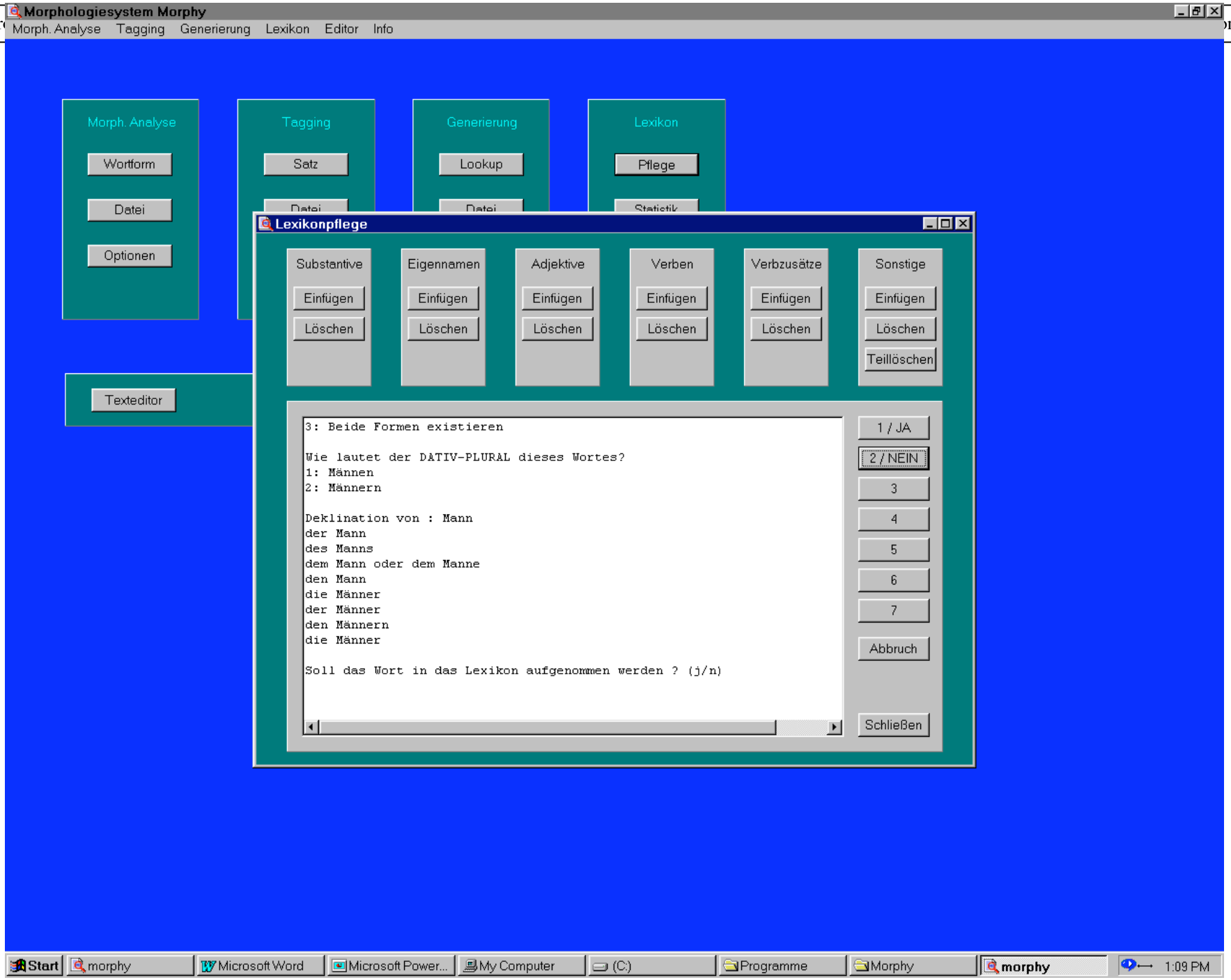

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
Eintrag des Substantivs: Mann
Handelt es sich um ein normales Substantiv (j=1) oder um eine Ausnahme
Wort ist bereits eingetragen. Ignorieren (j=1/n=2) ?
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:04 PM

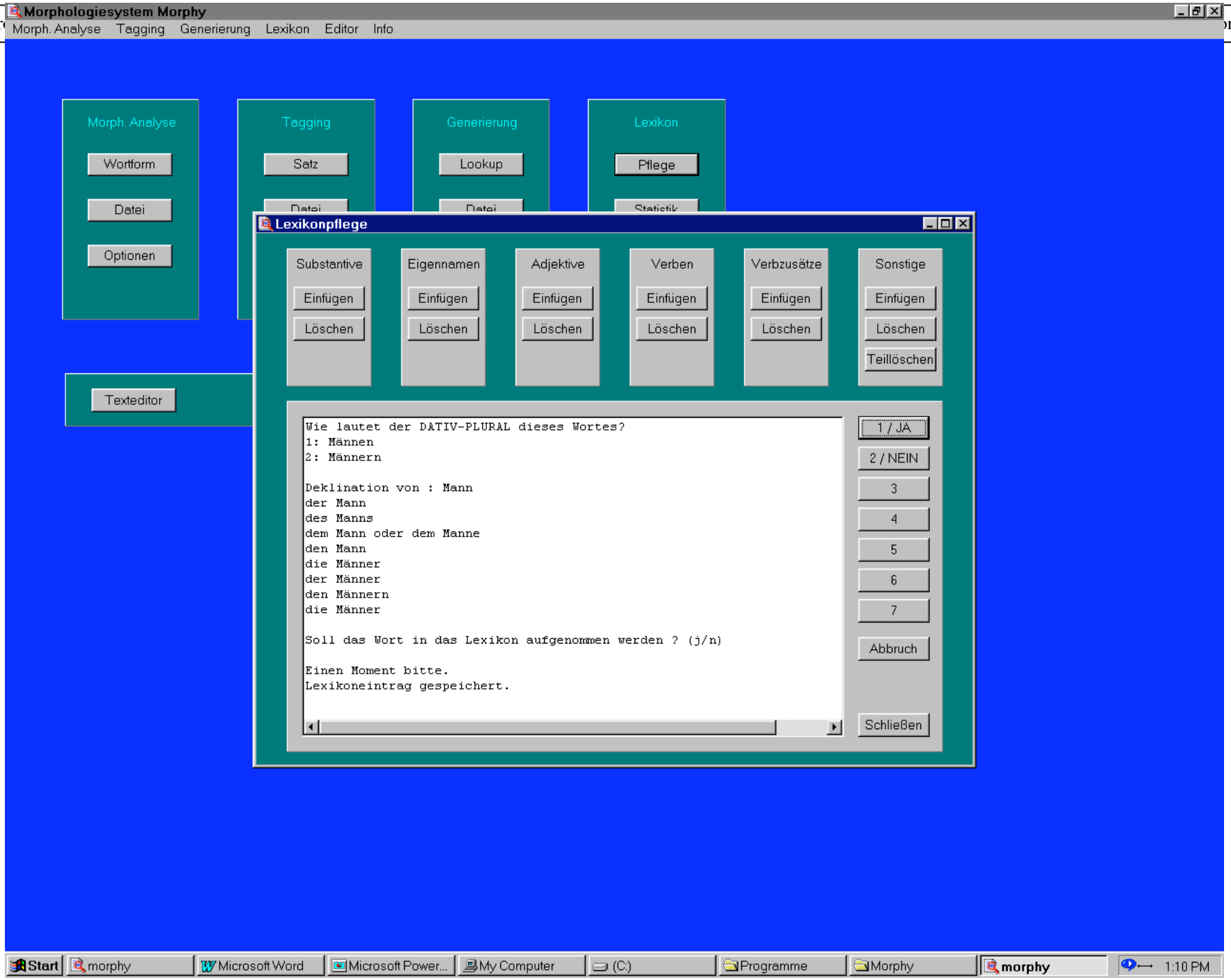

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
Eintrag des Substantivs: Mann
Handelt es sich um ein normales Substantiv (j=1) oder um eine Ausnahme
Wort ist bereits eingetragen. Ignorieren (j=1/n=2) ?
Welchen Genus hat dieses Wort ?
1: maskulin
2: feminin
3: neutrum
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:05 PM

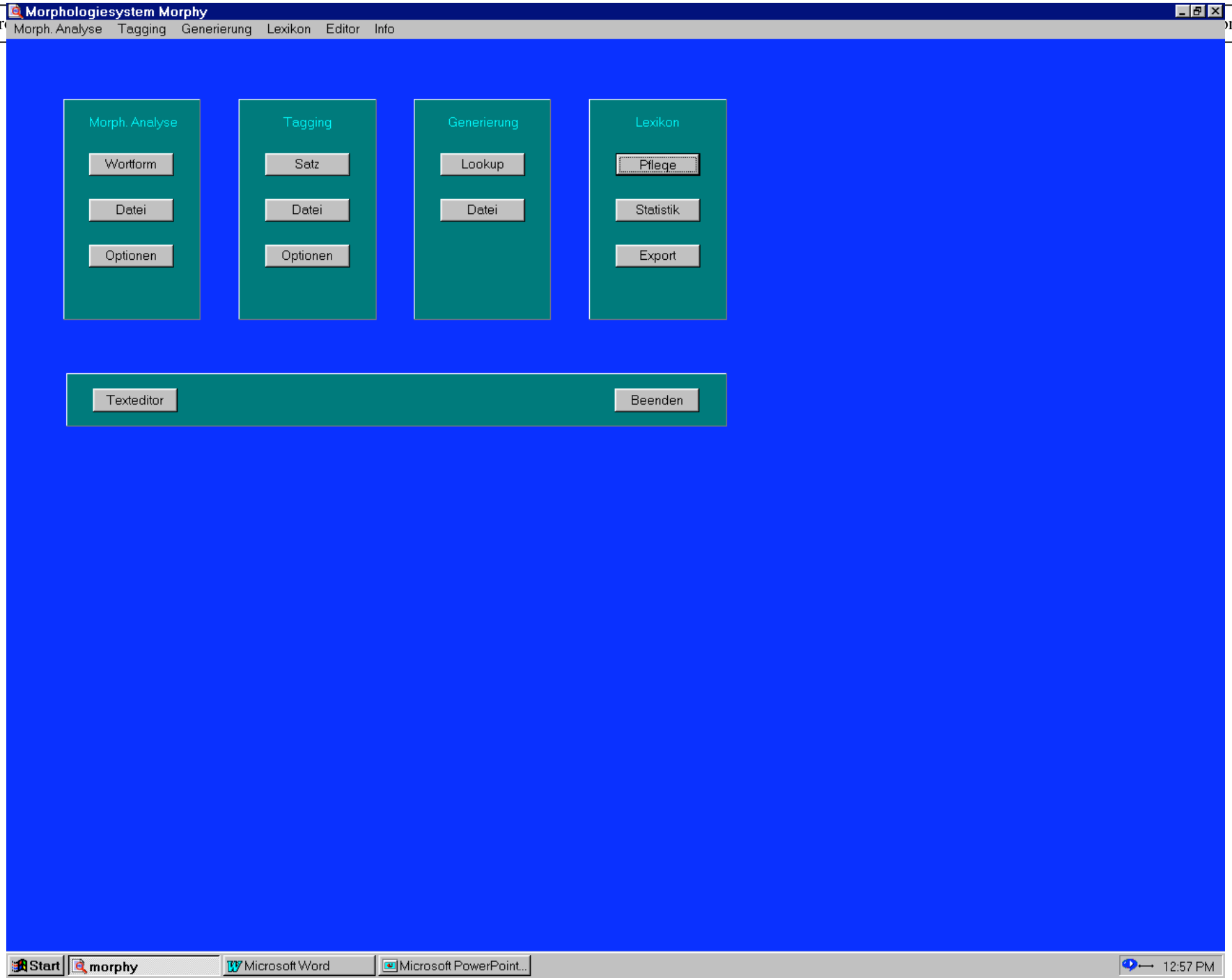

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
Eintrag des Substantivs: Mann
Handelt es sich um ein normales Substantiv (j=1) oder um eine Ausnahme
Wort ist bereits eingetragen. Ignorieren (j=1/n=2) ?
Welchen Genus hat dieses Wort ?
1: maskulin
2: feminin
3: neutrum
Muß bei der Pluralbildung umgelautet werden ?
1: Ja
2: Nein
3: Plural existiert nicht
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:06 PM

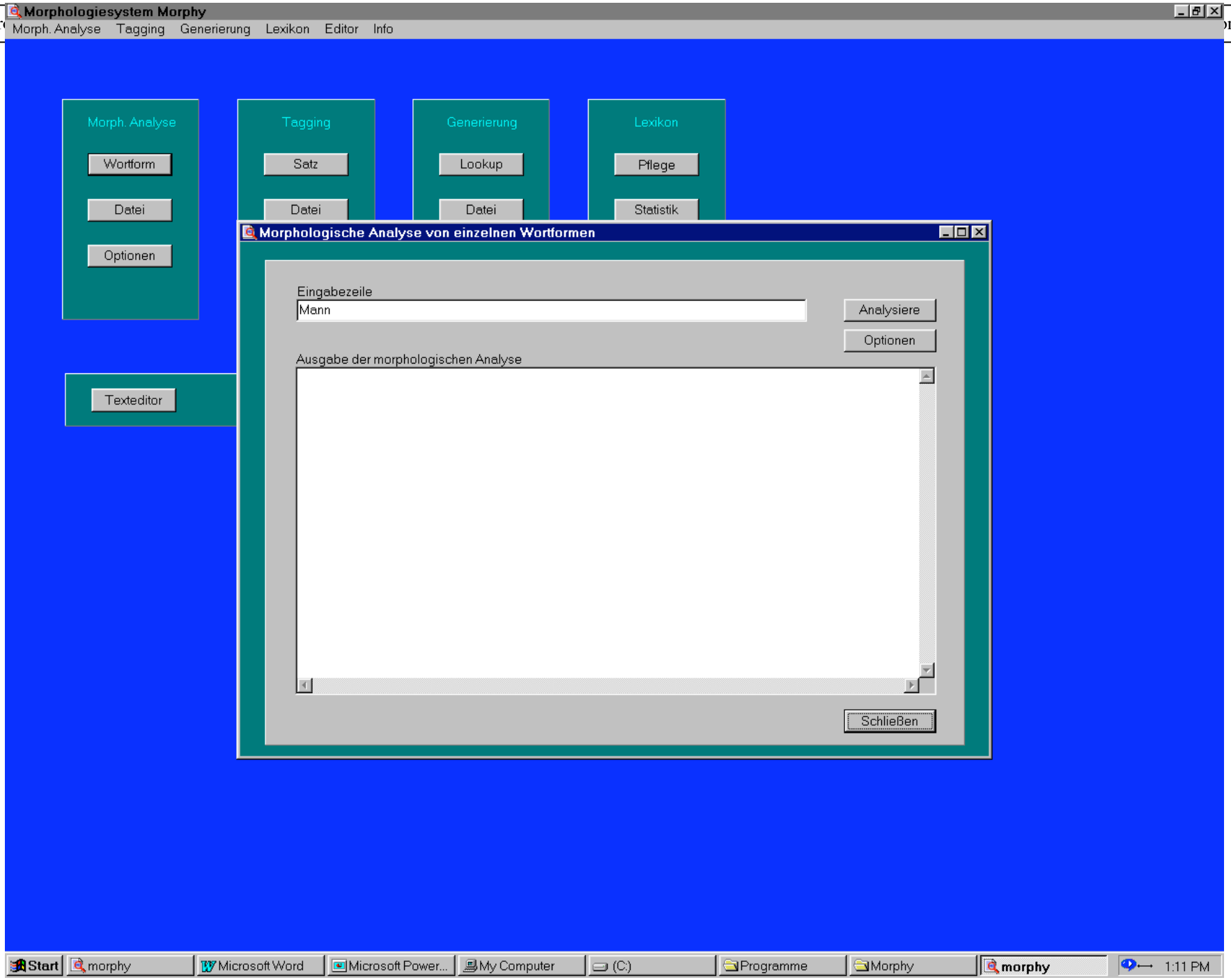

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
Wort ist bereits eingetragen. Ignorieren (j=1/n=2) ?
Welchen Genus hat dieses Wort ?
1: maskulin
2: feminin
3: neutrum
Muß bei der Pluralbildung umgelautet werden ?
1: Ja
2: Nein
3: Plural existiert nicht
Wie lautet der GENITIV-SINGULAR dieses Wortes?
1 : Manns
2 : Mannes/s
3 : Mannens
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:07 PM

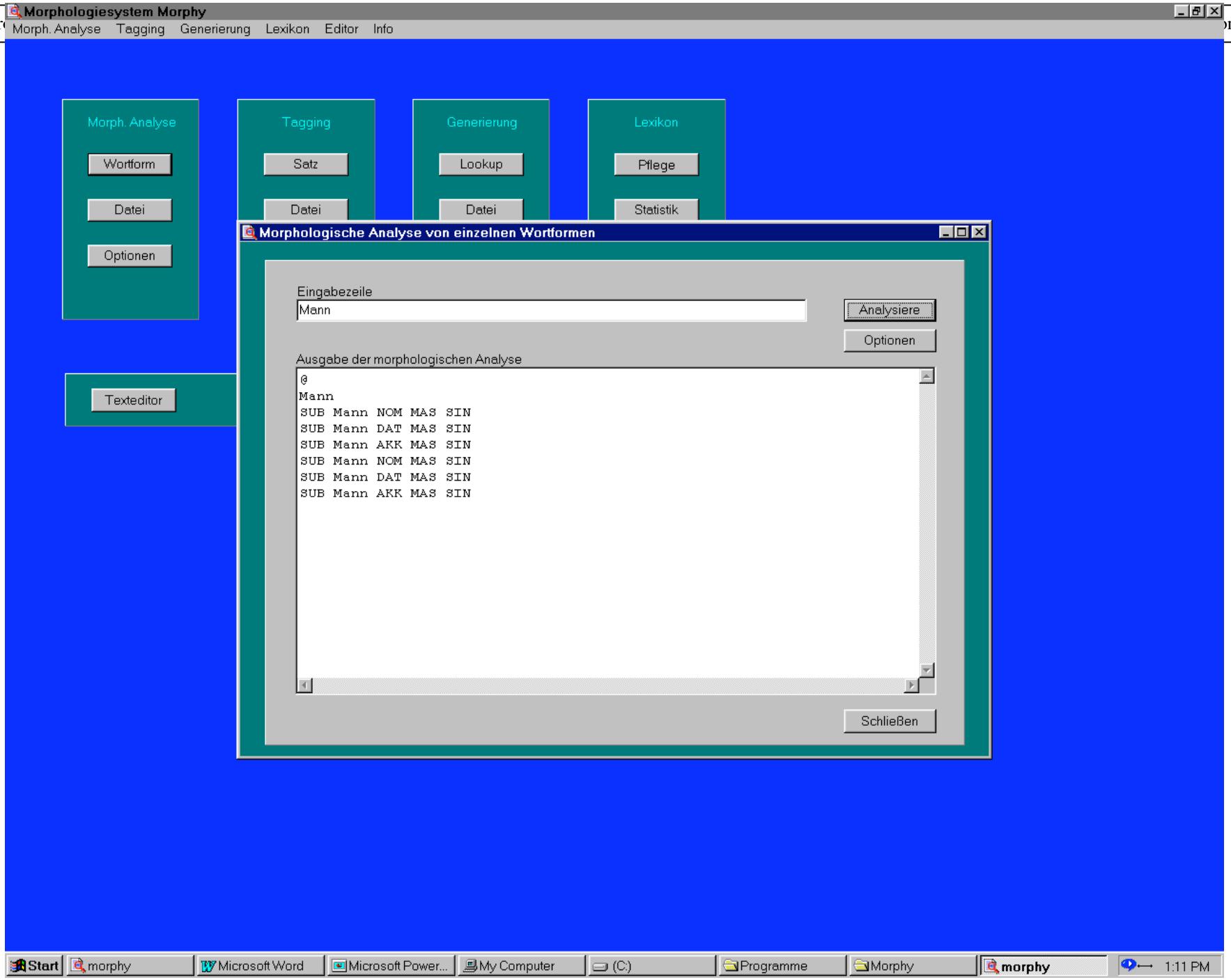

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
1: Ja
2: Nein
3: Plural existiert nicht
Wie lautet der GENITIV-SINGULAR dieses Wortes?
1 : Manns
2 : Mannes/s
3 : Mannens
Welche Formen existieren im GENITIV-SINGULAR ?
1: Mannes
2: Manns
3: Beide Formen existieren
Wie lautet der DATIV-PLURAL dieses Wortes?
1: Männen
2: Männern
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:08 PM

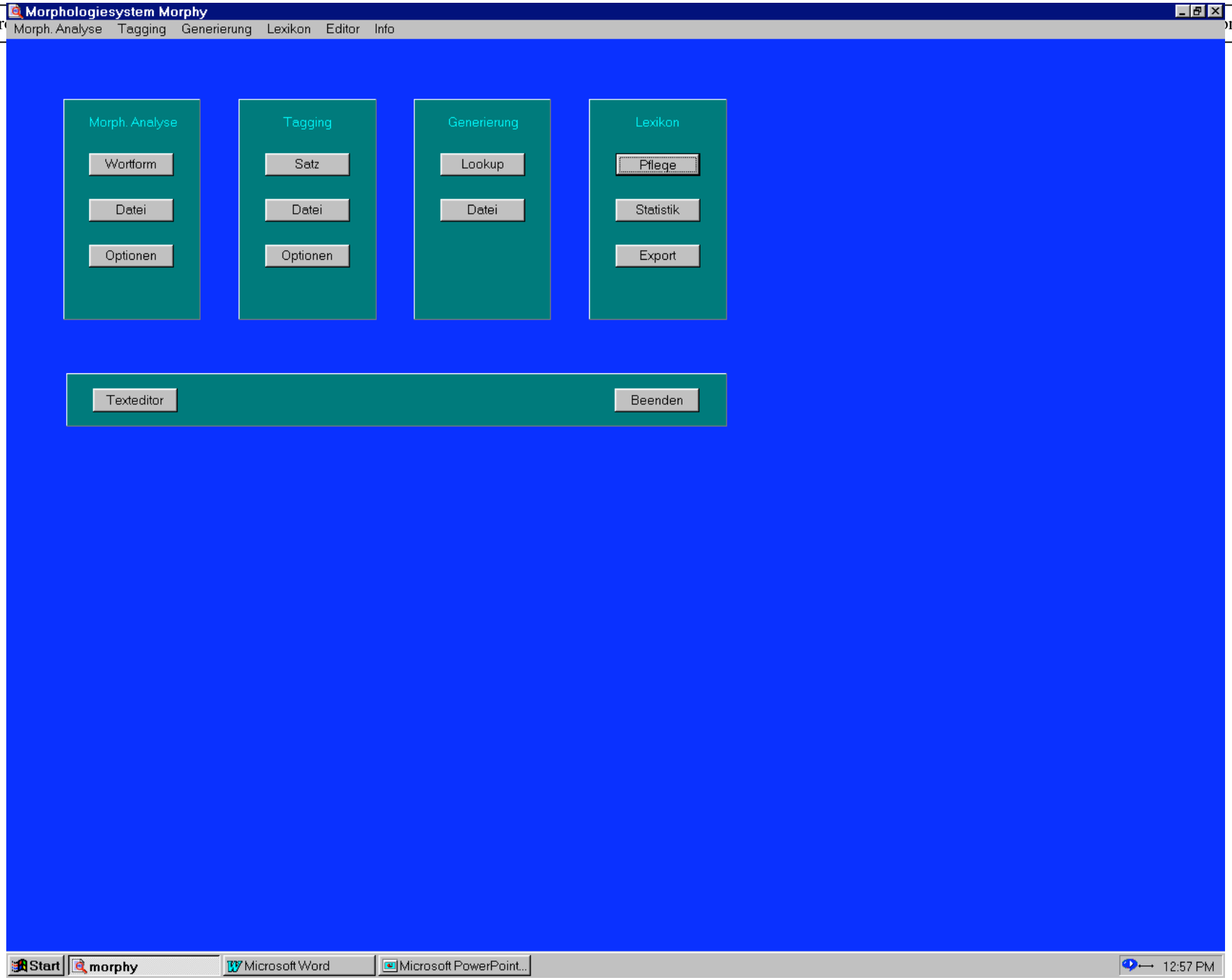

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
3: Beide Formen existieren
Wie lautet der DATIV-PLURAL dieses Wortes?
1: Männen
2: Männern
Deklination von : Mann
der Mann
des Manns
dem Mann oder dem Manne
den Mann
die Männer
der Männer
den Männern
die Männer
Soll das Wort in das Lexikon aufgenommen werden ? (j/n)
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:09 PM

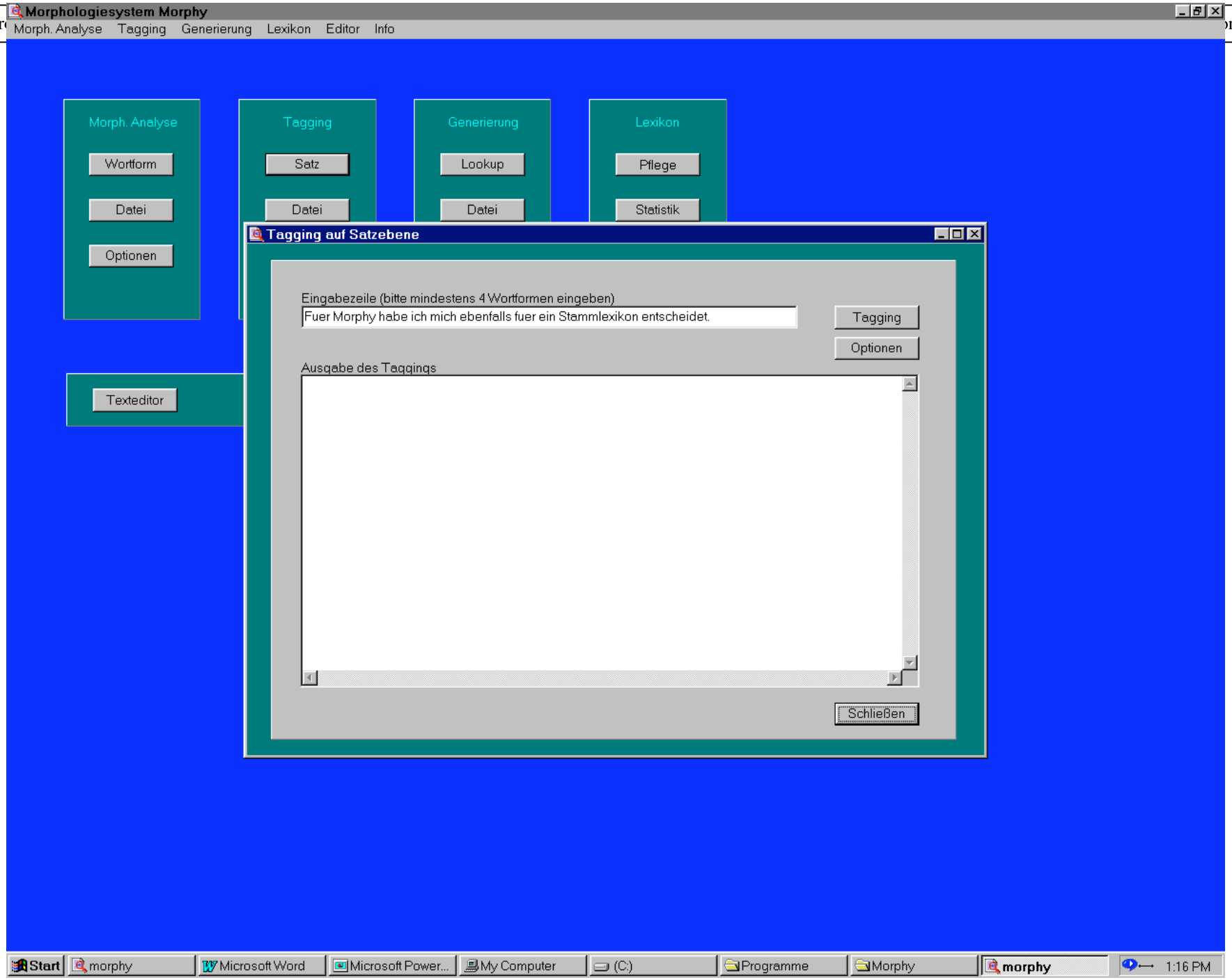

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Generierung
Lookup
Lexikon
Pflege
Texteditor
Lexikonpflege
Substantive
Einfügen
Löschen
Eigennamen
Einfügen
Löschen
Adjektive
Einfügen
Löschen
Verben
Einfügen
Löschen
Verbzusätze
Einfügen
Löschen
Sonstige
Einfügen
Löschen
Teillöschen
Wie lautet der DATIV-PLURAL dieses Wortes?
1: Männen
2: Männern
Deklination von : Mann
der Mann
des Manns
dem Mann oder dem Manne
den Mann
die Männer
der Männer
den Männern
die Männer
Soll das Wort in das Lexikon aufgenommen werden ? (j/n)
Einen Moment bitte.
Lexikoneintrag gespeichert.
1 / JA
2 / NEIN
3
4
5
6
7
Abbruch
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:10 PM

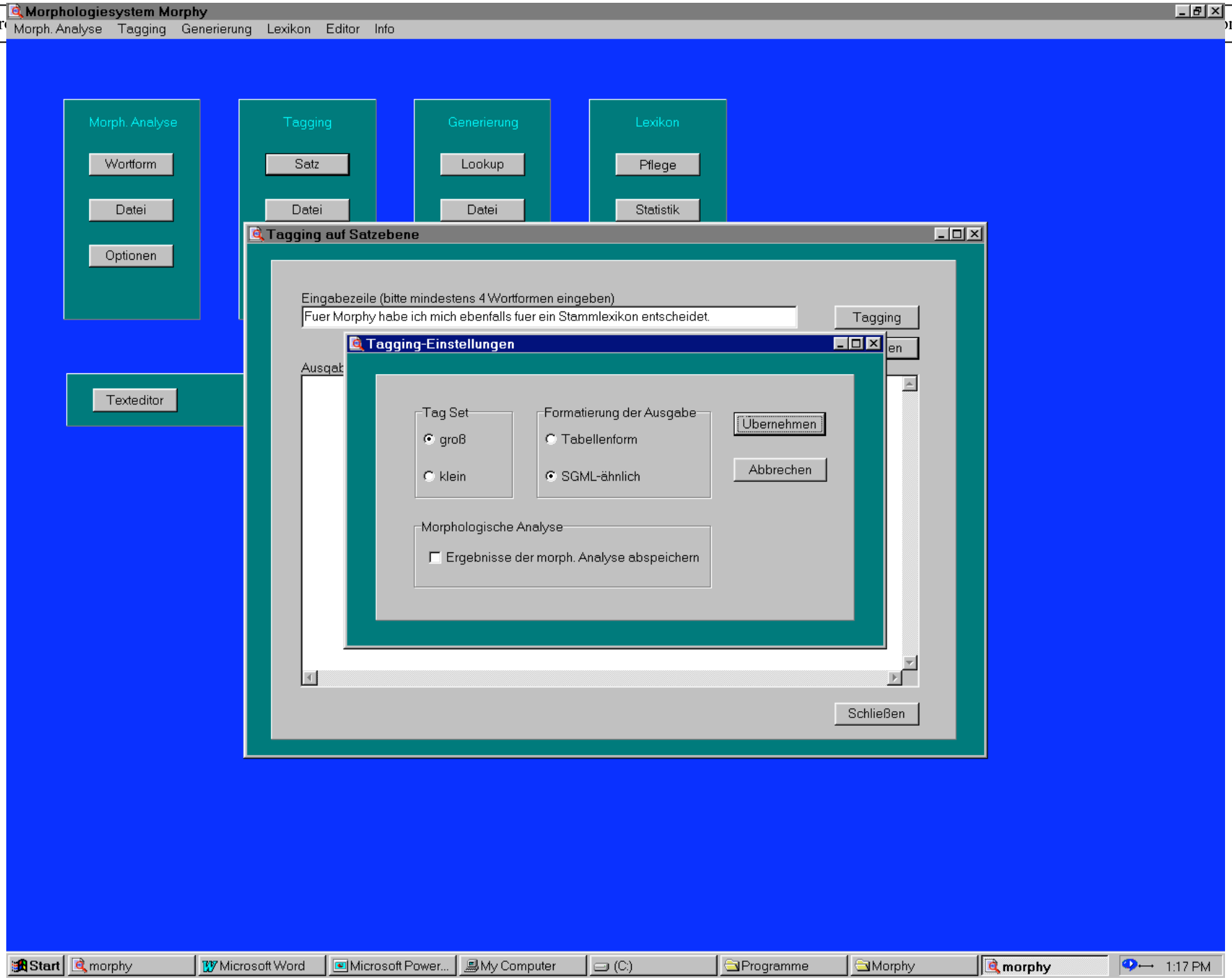

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Optionen
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Export
Texteditor
Beenden
Start
morphy
Microsoft Word
Microsoft PowerPoint...
12:57 PM

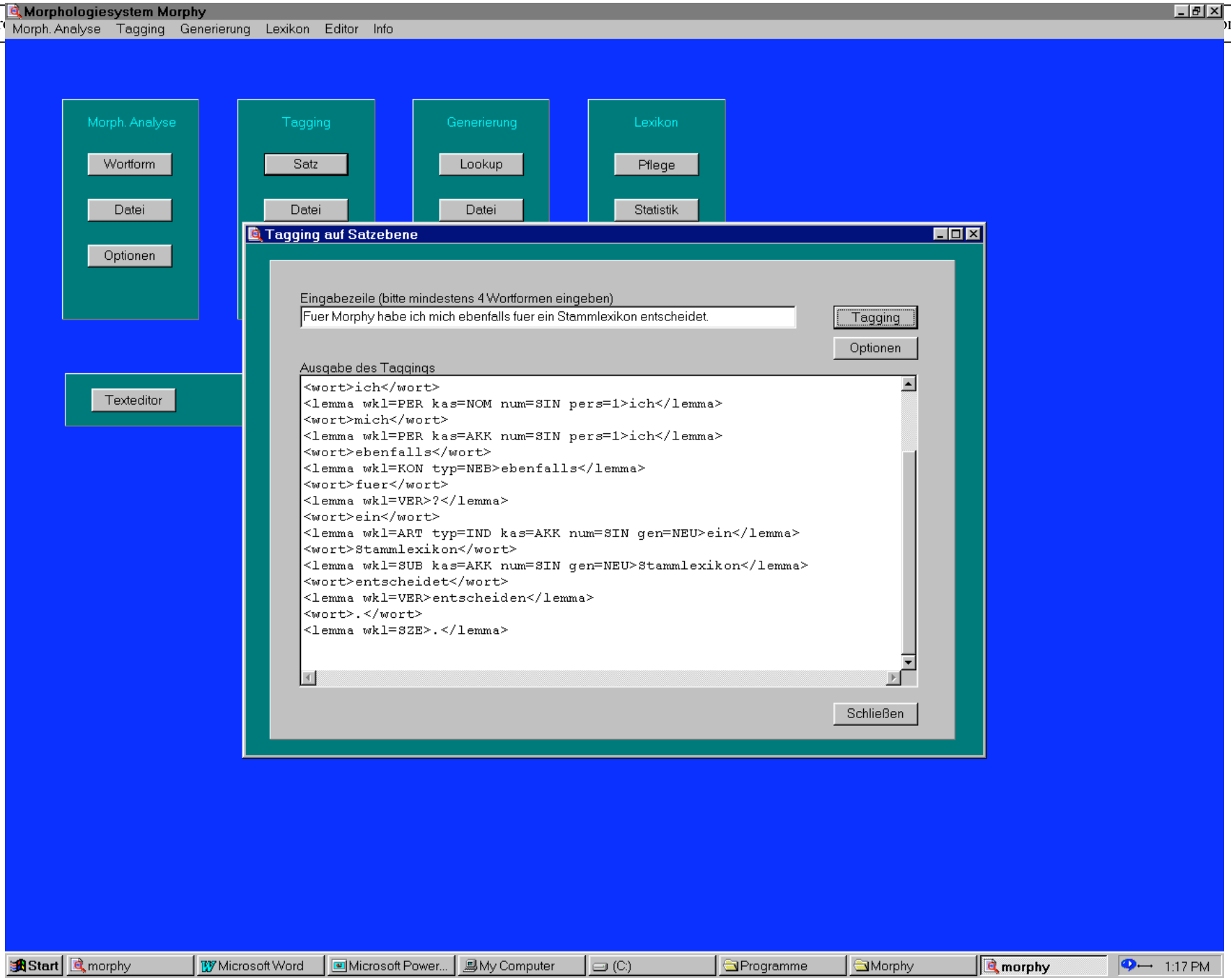

Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Texteditor
Morphologische Analyse von einzelnen Wortformen
Eingabezeile
Mann
Analysiere
Optionen
Ausgabe der morphologischen Analyse
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:11 PM


Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Texteditor
Morphologische Analyse von einzelnen Wortformen
Eingabezeile
Mann
Analysiere
Optionen
Ausgabe der morphologischen Analyse
@
Mann
SUB Mann NOM MAS SIN
SUB Mann DAT MAS SIN
SUB Mann AKK MAS SIN
SUB Mann NOM MAS SIN
SUB Mann DAT MAS SIN
SUB Mann AKK MAS SIN
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:11 PM


Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Optionen
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Export
Texteditor
Beenden
Start
morphy
Microsoft Word
Microsoft PowerPoint...
12:57 PM


Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Texteditor
Tagging auf Satzebene
Eingabezeile (bitte mindestens 4 Wortformen eingeben)
Fuer Morphy habe ich mich ebenfalls fuer ein Stammlexikon entscheidet.
Tagging
Optionen
Ausgabe des Taggings
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:16 PM


Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Texteditor
Tagging auf Satzebene
Eingabezeile (bitte mindestens 4 Wortformen eingeben)
Fuer Morphy habe ich mich ebenfalls fuer ein Stammlexikon entscheidet.
Tagging
Tagging-Einstellungen
Tag Set
groß
klein
Formatierung der Ausgabe
Tabellenform
SGML-ähnlich
Übernehmen
Abbrechen
Morphologische Analyse
Ergebnisse der morph. Analyse abspeichern
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:17 PM


Morphologiesystem Morphy
Morph. Analyse
Tagging
Generierung
Lexikon
Editor
Info
Morph. Analyse
Wortform
Datei
Optionen
Tagging
Satz
Datei
Generierung
Lookup
Datei
Lexikon
Pflege
Statistik
Texteditor
Tagging auf Satzebene
Eingabezeile (bitte mindestens 4 Wortformen eingeben)
Fuer Morphy habe ich mich ebenfalls fuer ein Stammlexikon entscheidet.
Tagging
Optionen
Ausgabe des Taggings
<wort>ich</wort>
<lemma wkl=PER kas=NOM num=SIN pers=1>ich</lemma>
<wort>mich</wort>
<lemma wkl=PER kas=AKK num=SIN pers=1>ich</lemma>
<wort>ebenfalls</wort>
<lemma wkl=KON typ=NEB>ebenfalls</lemma>
<wort>fuer</wort>
<lemma wkl=VER>?</lemma>
<wort>ein</wort>
<lemma wkl=ART typ=IND kas=AKK num=SIN gen=NEU>ein</lemma>
<wort>Stammlexikon</wort>
<lemma wkl=SUB kas=AKK num=SIN gen=NEU>Stammlexikon</lemma>
<wort>entscheidet</wort>
<lemma wkl=VER>entscheiden</lemma>
<wort>.</wort>
<lemma wkl=SZE>.</lemma>
Schließen
Start
morphy
Microsoft Word
Microsoft Power...
My Computer
(C:)
Programme
Morphy
morphy
1:17 PM

Lexikon als Entscheidungsbaum

- Sinn: effiziente Lexikonorganisation und -suche
- Kann erweitert werden durch wortartspezifische Fortsetzungsbäume, etwa für Nominalsuffixe, für Adjektivflexion etc.

Thesaurus-Basierte Desambiguierung: COOC

- Für die Bearbeitung von Texten ist es ein immer noch schwieriges Problem, die korrekten Lesarten von Lexemen bei jedem ihrer Vorkommen zu bestimmen. Beispiel: Ist bei dem Vorkommen von "Gericht" das Essen, das Gerichtsgebäude, die Institution oder die Rolle gemeint?
- In Seligman et. al. 99 ist ein Ansatz beschrieben, der mit Hilfe eines Thesaurus Zusammenhänge unterschiedlicher Stufen berechnen kann. Dabei wird davon ausgegangen, daß die Eingaben mit POS Tags annotiert sind.

- Projekt-Site:

  http://www.dfki.de/~janal/cooc.html
- Einführungsfolien unter:

  http://www.dfki.de/~janal/public_archive/cooc_eng.ppt

Lexikonorganisation

- Häufige Aufgabe für computerphilologische Arbeiten ist der Aufbau eines Lexikons und die Integration mehrerer Lexika.
- Über den Aufbau haben wir Alternativen gehört,
- Das Vereinigen (Merging) von Lexika ist eine komplexe Aufgabe:
  - Unterschiedliches Format:
    - *Leben: das, n, D23*
    - *<w pos='n'> leben < w form="Leben"></w>*
  - Unterschiedliche Kategorien:
    - *Partikeln*
    - *Funktionswörter*
  - Unterschiedliche Werte:
    - Kasus = {Nom, Gen, Dat, Acc, Voc, Abl}

Konkordanzen

- Konkordanzen sind Wort-Nachweise eines geschlossenen Texts (z.B. eines literarischen Texts oder Autors). Es gibt
  - Konkordanzen ohne Kontext (KWOC= key word out of context)
  - Konkordanzen mit Kontext (KWIC= key word in context)
- Probleme / Konzepte:
  - Werden alle Wörter aufgenommen oder nur „Inhaltswörter“
  - Wie lang ist der Kontext?
  - Wie werden Satzgrenzen behandelt?
  - Werden Wortformen lemmatisiert?
  - Werden Teilwörter gefunden?
  - Werden Homographen getrennt?

Thesauri

- Thesauri sind Wörterverzeichnisse einer Sprache in einer systematischen Anordnung
- Berühmt geworden sind
  - P. M. Rogets Thesaurus von 1852 und
  - für das Deutsche:
    - Dornseiff, Der Deutsche Wortschatz nach Sachgruppen und
    - Hugo Wehrle / Hans Eggers, Deutscher Wortschatz
- In loser Ausdrucksweise auch Wörterverzeichnisse anderer Anordnung.

Kollokationen, Idiome und Tokenizer

- Oft sind Wörter eines Textes im Sprachgebauch fest gebunden, d.h.
  - Idiome oder (gruppen-, zeit- oder regional-)gebundene Ausdrücke
  - Außerdem haben Wörter typische häufige Umgebungen, die man für Analysen verwenden oder berücksichtigen kann/muß (Kollokationen)
  - Im Deutschen gibt es Verbpartikel in Fernstellung („... hob das Urteil gegen .... auf“
- Die lexikalische Einheit ist daher manchmal nicht das Wort, sondern eine Menge von Wörtern (ling.: Mehrwort-Lexem).
- Um einen Text zunächst in die lexikalisch-semantischen Einheiten zu zerlegen, benutzt man Tokenizer, die aus einem Korpus durch Vergleich die Festigkeit einer Kollokation prüfen.

WordNet

- WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory.
- English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

- WordNet was developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller (Principal Investigator). Over the years, many people have contributed to the success of WordNet.

- http://www.cogsci.princeton.edu/

Deutsches WordNet

GermaNet - Introduction

GermaNet is a lexical-semantic net that has been developed within the LSD Project at the Division of Computational Linguistics of the Linguistics Department at the University of Tübingen. Currently it is being integrated into the EuroWordNet (EWN), a multilingual lexical-semantic database.

GermaNet relates German nouns, verbs, and adjectives semantically by grouping words belonging to the same concept and by defining semantic relations between concepts. It has much in common with the English WordNet® and might be viewed as an on-line thesaurus defining an explicit ontology.

If you want to get more information about GermaNet you might like to read the following paper:

Kunze, Claudia and Andreas Wagner (1999): Integrating GermaNet into EuroWordNet, a multilingual lexical-semantic database. In: "Sprache und Datenverarbeitung - International Journal for Language Data Processing". Vol. 23.2/1999Bonn, 1999.

http://www.sfs.nphil.uni-tuebingen.de/lsd/

Abteilung Automatische Sprachverarbeitung

Aktuelles

- Dr. Christian Wolff: *Sichere Kommunikation im Internet mit Hilfe angewandter Kryptographie*
  6. 12. 2000, 14 Uhr c.t., Raum SG 00-90
  Lehrprobe im Rahmen des Habilitationsverfahrens
- Am 1. und 2. Frebruar 2001 findet das Symposium TAMA 2001 - Sharing Terminological Knowledge - in Antwerpen statt.
- Folien (PPT) zur Vorlesung *Textfassung: Grundlagen und Anwendungen* von Professor Dr. G. Heyer im Rahmen der Ringvorlesung des Graduiertenkollegs Wissensrepräsentation
- Das Diplomanden- und Abteilungsseminar der ASV findet dienstags 11 Uhr im Raum HG 1-74 statt.

Mitarbeiter:

Leiter der Abteilung:
Prof. Dr. Gerhard Heyer, Studiendekan, Zi. 1-53, Tel. 32231

Sekretärin:
Renate Schildt, Zi. 1-52, Tel. 32230

Wissenschaftliche Mitarbeiter:
Dr. habil. Uwe Quasthoff, Zi. 1-50, Tel. 32233
Thomas Wittig, Zi. 2-21, Tel. 32232
Dr. habil. Christian Wolff, Zi. 1-51, Tel. 32249

Programmiererin:
Regine Gabler, Zi. 2-21, Tel. 32232

Kollegiat:
Dr. Martin Läuter, Zi. 1-42, Tel. 32302

Adresse

Abteilung Automatische Sprachverarbeitung
Institut für Informatik
Universität Leipzig
Augustusplatz 10-11
04109 Leipzig

Lexikalische Graphen - 1 -

http://wortschatz.informatik.uni-leipzig.de/wort/inhalt.htm

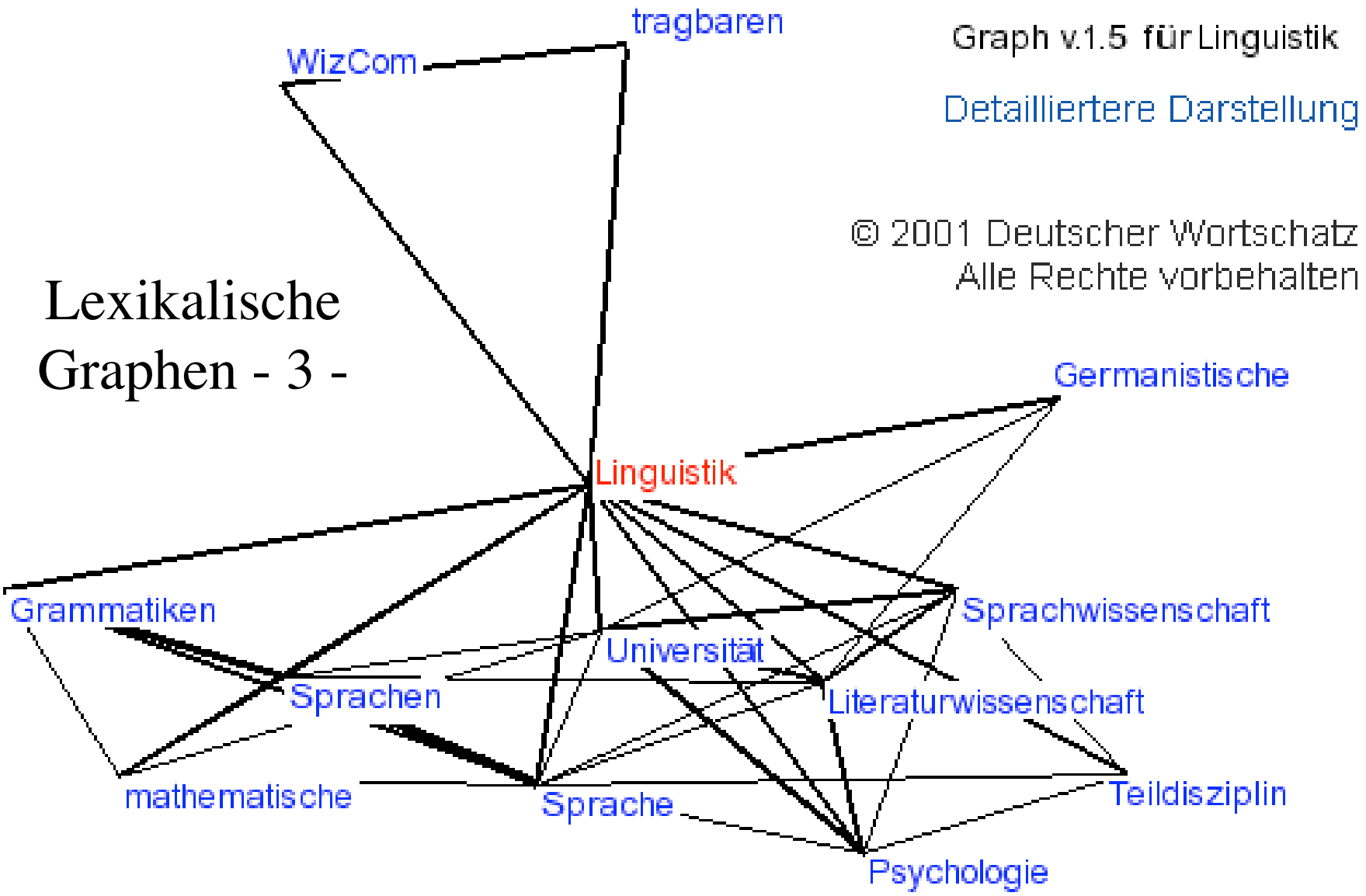
Wortschatz : Sitemap

Überblick über alle Seiten des Projekts:

- Suchoptionen im Projekt Deutscher Wortschatz
  - einfache Suche
  - erweiterte Suche
    - Abfrage von Anagrammen
    - Suche Adjektiv zu Substantiv
    - Suche Verb zu Substantiv
    - Wörter bestimmter Länge
- Downloads zum Projekt Deutscher Wortschatz:
  - Dokumente über das Projekt
  - Das Programm Satzsegmentierer
  - Wortlisten mit den häuftigsten deutschen und englischen Wörtern
- Informationen zum Projekt Deutscher Wortschatz:
  - Mitarbeiter des Projekts
  - Partnerlexika des Lexikon-Portals
  - Partner des Projekts
  - Die Wortschatz CD-ROM
  - Kontakt zum Projekt
  - Links zu verwandten Projekten
  - Ihre Meinung ist gefragt: Umfragen
  - Was uns Besucher schrieben...

Lexikalische Graphen - 2 -

Lexikalische Graphen - 3 -



Lexikalische Graphen - 4 -

Wortschatz : Suche : **Ergebnis**

Zum Haupteintrag Linguistik

Signifikante Kollokationen für Linguistik:

Sprache (50), mathematische (42), Teildisziplin (33), Psychologie (31), Sprachwissenschaft (27), Sprachen (26), Grammatiken (25), Literaturwissenschaft (24), Germanistische (23), Saussure (23), tragbaren (23), Universität (20), WizCom (20), Semiotik (19), Soziologie (19), Professor (18), Soziolinguistik (17), Sprachstruktur (17), Noam Chomsky (16), deduktiv (16), komparative (16), sprachwissenschaftlichen (16), Phonetik (15), Phonologie (15), sprachlicher (15), feministische (14), Reduktionsalgorithmen (13), Teilgebiet (13), Truncation-Verfahren (13), linguistischen (13), synchronischen (13), Deborah Tannen (12), Entwicklung (12), Informatik (12), Institut (12), Linguistik (12), Linguistik (12), groupes (12), natürlichen (12), nominaux (12), Philologie (11), Philosophie (11), Poetik (11), Psycholinguistik (11), Quicktionary (11), Transformationsgrammatik (11), bedeutungstragende (11), Computerlinguistik (10), Syntax (10), langue (10), Ammon (9), Cours (9), Pragmatik (9), Professorin (9), Religionswissenschaft (9), Scannern (9), Völkerwanderungen (9), Wissenschaft (9), abstrahiert (9), allgemeinen (9), forensische (9), modernen (9), sprachliche (9), Bedeutung (8), Kategorien (8), beschäftigt (8), erforscht (8), kognitiven (8), untersucht (8), Anthropologie (7), Medium (7), Methodik (7), Struktur (7), Untersuchung (7), auseinandersetzt (7), empirische (7), führender (7), interdisziplinäre (7), vergleichenden (7), weltweit (7)

Signifikante linke Nachbam von Linguistik:

Germanistische (26), feministische (18), komparative (18), synchronischen (14), modernen (11), forensische (10), klassische (6), Bereichen (5), historische (5), historischen (5)

Signifikante rechte Nachbam von Linguistik:

beschäftigt (7)

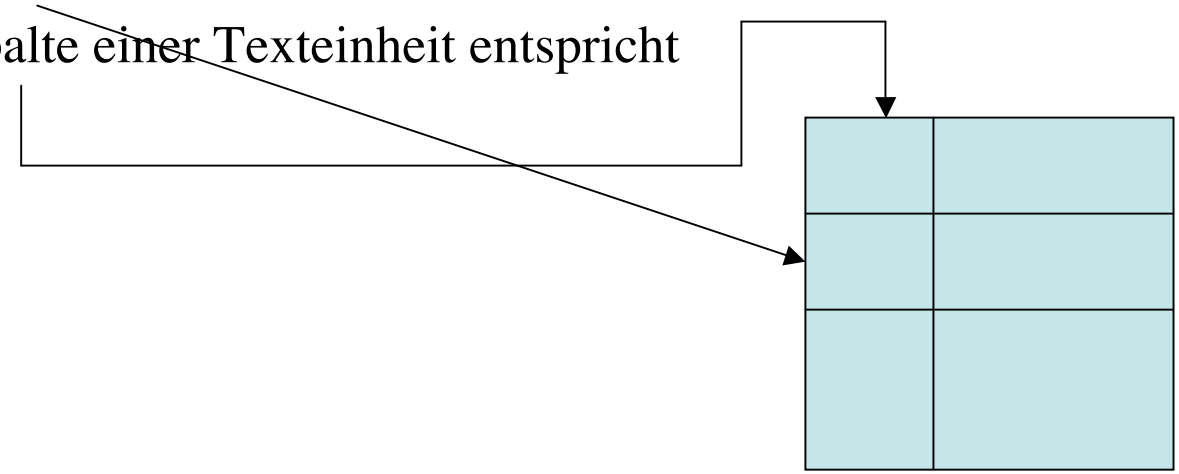
Latent Semantic Analysis (LSA)

- LSA ist eine vollautomatische statistische Methode, um aus sehr großen Textmengen die Wahrscheinlichkeit von lexikalisch-semantischen Beziehungen (Ähnlichkeiten) zu erheben und in großen Matrizen (ca 100 x 500) darzustellen.
- LSA arbeitet ohne:
  - Lexikon
  - Wissensbasis
  - Semantische Netze
  - Syntaktische Parser
  - Morphologie

- Home-Page: http://lsa.colorado.edu/

Latent Semantic Analysis
-Eingabe-

- Die LSA-Eingabe ist allein der Rohtext:
  - in Wörter segmentiert (ein Wort = ein einziger String)
  - in bedeutungsvolle Passagen getrennt (Sätze, Paragraphen)
- Der Text wird in eine Matrix eingelesen, in der:
  - jede Zeile einem Wort (einem type) und
  - jede Spalte einer Texteinheit entspricht



Latent Semantic Analysis
-Texteingabe-Beispiel -

Benutzt werden Wörter, die in mindestens 2 Titeln erscheinen (außer extrem häufige Funktionswörter)

Text (Titel technischer Berichte):

c1: *Human* machine *interface* for *computer* applications

c2: A *survey* of *user* opinion of *computer system response time*

c3: The *EPS user interface* management *system*

c4: System and *human system engineering testing of EPS*

c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*

m2: The intersection *graph* of paths in *trees*

m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering

m4: *Graph minors*: A *survey*

Spaltenanordnung

m4: *Graph minors*: A *survey*

m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering

m2: The intersection *graph* of paths in *trees*

m1: The generation of random, binary, ordered *trees*

c5: Relation of *user* perceived *response time* to error measurement

c4: System and *human system engineering testing of EPS*

c3: The *EPS user interface* management *system*

c2: A *survey* of *user* opinion of *computer system response time*

c1: *Human* machine *interface* for *computer* applications

Latent Semantic Analysis
-Textmatrizen-Beispiel -

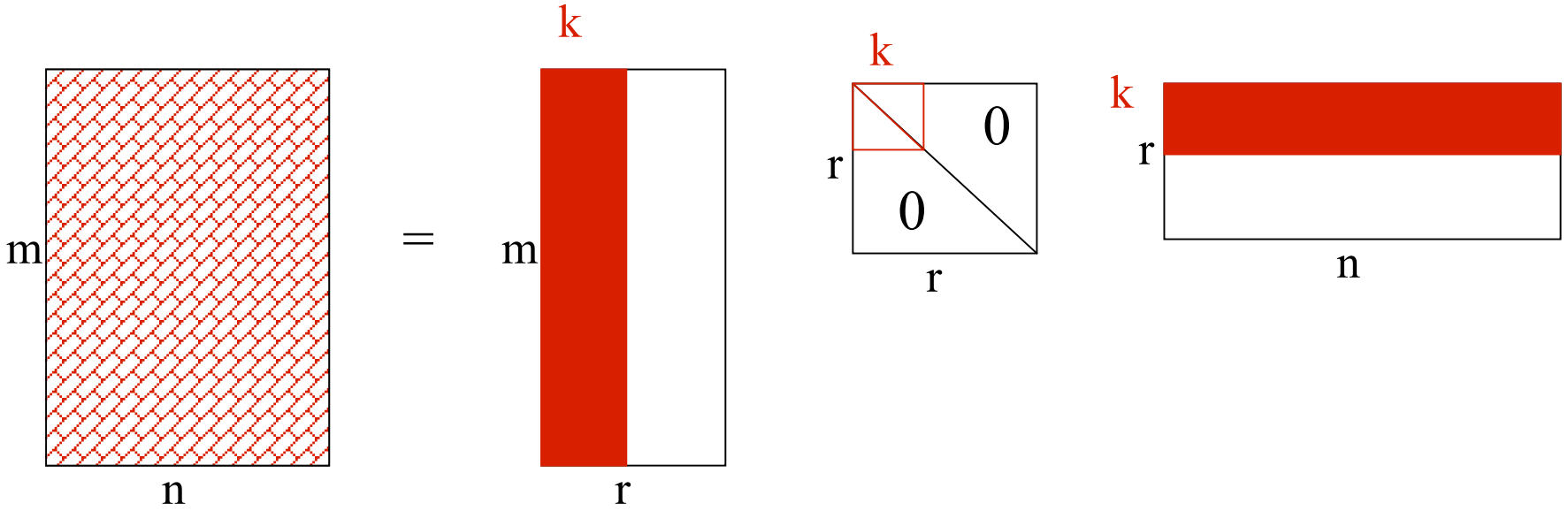
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	1	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Auftretenshäufigkeit im entspechenden Satz

In realistischen Anwendungen werden die Zellen gewichtet.

Latent Semantic Analysis
-Matrizentransformation -

SVD - Single Value Decomposition



A = B × I × C

Je größer die Matrizen umso größer der Berechnungsaufwand

A_k = B_k × I_k × C_k

Latent Semantic Analysis
- Matrizentransformations-Beispiel -

Matrizenrekonstruktion mit k=2

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

"tree" erscheint nicht in m4, aber in Titeln mit "graph" und "minors"

korr(human,user) = 0.94

korr(human,minors) = -0.83

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

korr(human,user) = 0.38

korr(human,minors) = 0.29

c1	c2	c3	c4	c5	m1	m2	m3	m4
1	0	0	1	0	0	0	0	0
1	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0
0	1	1	0	1	0	0	0	0
0	1	1	2	0	0	0	0	0
0	1	0	0	1	0	0	0	0
0	1	0	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0
0	1	0	1	0	0	0	0	1
0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	1

Original

LSA als Modell menschlichen konzeptuellen Wissens (LSA Evaluationsliste)

- predictor of query ⇔ document topic similarity judgements
- a simulation of agreed upon word ⇔ word relations and of human vocabulary test synonym judgements
- a simulation of human choices on subject-matter multiple choice tests
- a predictor of text coherence and resulting comprehension
- a simulation of word ⇔ word and passage ⇔ word relations found in lexical priming experiments
- subjective ratings of text properties (i.e. grades assigned to essays).
- a predictor of appropriate matches of instructional text to learners,
- to mimic synonym, antonym, singular-plural and compound word relations