(19) Bundesrepublik Deutschland
Deutsches Patent- und Markenamt

(10) **DE 693 33 568 T2** 2004.10.21

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 0 658 259 B1**
(21) Deutsches Aktenzeichen: **693 33 568.8**
(86) PCT-Aktenzeichen: **PCT/US93/07928**
(96) Europäisches Aktenzeichen: **93 920 279.2**
(87) PCT-Veröffentlichungs-Nr.: **WO 94/06086**
(86) PCT-Anmeldetag: **23.08.1993**
(87) Veröffentlichungstag der PCT-Anmeldung: **17.03.1994**
(97) Erstveröffentlichung durch das EPA: **21.06.1995**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **01.03.2000**
(47) Veröffentlichungstag im Patentblatt: **21.10.2004**

(51) Int Cl.[7]: **G06F 17/27**
G06F 17/28

(30) Unionspriorität:
941180 **04.09.1992** **US**

(73) Patentinhaber:
Caterpillar Inc., Peoria, Ill., US

(74) Vertreter:
WAGNER & GEYER Partnerschaft Patent- und Rechtsanwälte, 80538 München

(84) Benannte Vertragsstaaten:
AT, BE, CH, DE, ES, FR, GB, IT, LI, NL, SE

(72) Erfinder:
CARBONELL, Jaime G., Pittsburgh, PA 15217, US; GALLUP, Sharlene L., Morton, IL 61550, US; HARRIS, Timothy J., Pekin, IL 61554, US; HIGDON, James W., Lacon, IL 61540, US; HILL, Dennis A., East Peoria, IL 61611, US; HUDSON, David C., Edelstein, IL 61526, US; NASJLETI, David, Morton, IL 61550, US; RENNICH, Mervin L., Dunlap, IL 61525, US; ANDERSON, Peggy M., Pittsburgh, PA 15217, US; BAUER, Michael M., Pittsburgh, PA 15232, US; BUSDIECKER, Roy F., III, Pittsburgh, PA 15232, US; HAYES, Philip J., Pittsburgh, PA 15217, US; HUETTNER, Alison K., Pittsburgh, PA 15206, US; McLAREN, Bruce M., Pittsburgh, PA 15206, US; NIRENBURG, Irene, Pittsburgh, PA 15217, US; RIEBLING, Eric H., Pittsburgh, PA 15237, US; SCHMANDT, Linda M., Pittsburgh, PA 15217, US; SWEET, John F., Pittsburgh, PA 15206, US; BAKER, Kathryn L., Pittsburgh, PA 15206, US; BROWNLOW, Nicolas D., Pittsburgh, PA 15217, US; FRANZ, Alexander M., Pittsburgh, PA 15217, US; HOLM, Susan E., Pittsburgh, PA 15217, US; LEAVITT, John Robert Russell, Pittsburgh, PA 15232, US; LONSDALE, Deryle W., Bridgeville, PA 15017, US; MITAMURA, Teruko, Pittsburgh, PA 15213, US; NYBERG, Eric H., III, Pittsburgh, PA 15213, US

(54) Bezeichnung: **INTEGRIERTES ENTWURF- UND ÜBERSETZUNGSSYSTEM.**

Beschreibung

Hintergrund der Erfindung

1. Gebiet der Erfindung

[0001] Die vorliegende Erfindung bezieht sich im allgemeinen auf computerbasierende Dokumentenerzeugung und auf ein Übersetzungssystem und, insbesondere, auf ein System zum Verfassen bzw. Entwerfen und Übersetzen von sprachlich eingegrenzten Texten in eine Fremdsprache ohne das Erfordernis von Vor- oder Nacheditierung.

2. Stand der Technik

[0002] Jede Organisation, deren Aktivitäten die Erzeugung einer großen Menge von Information in einer Vielzahl bzw. Verschiedenartigkeit von Dokumenten erfordern, ist mit der Notwendigkeit konfrontiert, die Lesbarkeit der zuvor genannten Dokumente sicherzustellen. Idealer Weise, sollten solche Dokumente in einer einfachen, direkten Sprache verfaßt sein, die alle notwendigen Ausdrucksattribute für eine optimierte Kommunikation zeigt. Diese Sprache sollte konsistent sein, so daß die Organisation durch ihre einzige, stabile Stimme identifiziert wird. Diese Sprache sollte unzweideutig sein. Das Streben nach dieser Art von Schreibfähigkeit führte zu Ausführungen von verschiedenen Disziplinen, die ausgelegt sind, den Prozeß des Verfassens bzw. Entwerfens unter Kontrolle zu bringen. Dennoch können Autoren bzw. Verfasser mit verschiedenen Fähigkeiten und Hintergründen nicht einfach bzw. beschwerdefrei dazu gebracht werden, sich in einem einheitlichen Fertigkeitsstandard einzupassen. Richtlinien, Regeln und Standards für das Schreiben sind schwer zu begreifen – schwierig zu definieren und durchzusetzen. Bemühungen, die auf die Standardisierung und die Verbesserung der Qualität des Schreibens gerichtet sind, haben die Tendenz auf gemischte Resultate zu stoßen. Wenn sie jedoch erreicht und jedoch erfolgreich sind, treiben diese Resultate die Kosten der Dokumentenverfassung hoch.

[0003] Jüngste Versuche, Autoren bzw. Verfasser in eine Softwareumgebung einzubinden, die ihre Produktivität und die Qualität ihres Schreibens erhöhen könnte, resultierten nur im Vorsehen von Buchstabierüberprüfungen. Die Effektivität von anderer Schreibsoftware war bisher enttäuschend schwach.

[0004] Wenn der Lieferbedarf nach Information nach dem Überschreiten von linguistischen Grenzen ruft, vervielfachen sich die Herausforderungen. Die Organisation, die einen Kanal für ihren Informationsfluß freigeben muß, findet sich selbst in einem großen Ausmaß, wenn nicht sogar absolut, abhängig von Übersetzungen.

[0005] Textübersetzungen von einer Sprache in eine andere Sprache wurden seit hunderten von Jahren durchgeführt. Vor dem Aufkommen von Computern wurden solche Übersetzungen völlig manuell durch Experten, sogenannte Übersetzer, durchgeführt, die flüssig in der Sprache des Originaltextes (Quellentext) und in der Sprache des übersetzten Textes (Zieltext) waren. Typischer Weise war es bevorzugt, daß der Übersetzer ursprünglich die Zielsprache als seine/ihre Muttersprache lernte und nachfolgend die Quellensprache lernte. Solch eine Herangehensweise wurde so betrachtet, daß sie die genaueste und effizienteste Übersetzung lieferte.

[0006] Selbst der fachkundigste Übersetzer beansprucht eine beträchtliche Zeitdauer zur Übersetzung einer Textseite. Beispielsweise wird geschätzt, daß ein fachkundiger Übersetzer bei der Übersetzung eines technischen Textes von Englisch in Japanisch nur ungefähr 300 Worte (ungefähr eine Seite) pro Stunde übersetzen kann. Daraus ist ersichtlich, daß die Zeitdauer und der Aufwand, der zur Übersetzung eines Dokuments nötig ist, insbesondere eines technischen, sehr extensiv ist.

[0007] Die Erfordernisse für eine Übersetzung im Geschäftsleben und Handel wuchs stetig in den letzten hundert Jahren an. Dies beruht auf mehreren Faktoren. Einer ist das schnellen Anwachsen bei Texten, die assoziiert mit der Durchführung mit internationalen Geschäften sind. Ein weiterer ist die große Anzahl von Sprachen, in die solche Texte übersetzt werden müssen, damit eine Firma sich im globalen Handel engagieren bzw. einsetzen kann. Ein dritter ist der schnelle Fortschritts des Handels, was in einer häufigen Überarbeitung der Textdokumente resultierte, was wiederum eine nachfolgende Übersetzung der neuen Versionen erforderte.

[0008] Viele Organisationen tragen die Verantwortung der Erzeugung und Verteilung von Information in einer Vielzahl von Sprachen. Im globalen Markt muß der Hersteller sicherstellen, daß die Betriebsanleitungen bzw. Manuals weit verbreitet verfügbar in den Gastsprachen ihrer Zielmärkte sind. Manuelle Übersetzung von Do-

kumenten in Fremdsprachen ist ein teurer, zeitaufwendiger und ineffizienter Prozeß. Übersetzungen sind für gewöhnlich inkonsistent, und zwar aufgrund der individuellen Interpretation der Übersetzer, die nicht notwendiger Weise gut versiert in der Anwendung der spezifischen Sprache sind, die in der Dokumentation verwendet wird. Aufgrund dieser Probleme werden weniger Anleitungen tatsächlich übersetzt als idealer Weise der Fall wäre.

[0009] Auf den Gebieten der Forschung und Entwicklung hat die Wissensexplosion, die im letzten Jahrhundert eintrat, ebenso geometrisch den Bedarf für die Übersetzung von Dokumenten erhöht. Es gibt nicht länger eine dominierende Sprache für Dokumente in einem besonderen Gebiet der Forschung und Entwicklung. Typischer Weise finden solche Forschungs- und Entwicklungsaktivitäten in mehreren, fortgeschritten industrialisierten Ländern statt, wie beispielsweise den Vereinigten Staaten, Großbritannien, Frankreich, Deutschland und Japan. Oftmals gibt es zusätzliche Sprachen, die wichtige Dokumente umfassen, die sich auf ein besonderes Gebiet der Forschung und Entwicklung beziehen. Fortschritte in der Technologie, insbesondere der Elektronik und Computer, haben weiter die Produktion von Texten in allen Sprachen beschleunigt.

[0010] Die Möglichkeit der Erzeugung von Text ist direkt proportional zur Fähigkeit der verwendeten Technologie. Als Dokumente handgeschrieben werden mußten, konnte beispielsweise ein Verfasser nur eine bestimmte Anzahl von Worten pro Zeiteinheit erzeugen. Dies stieg signifikant jedoch mit dem Aufkommen von mechanischen Geräten an, wie beispielsweise Schreibmaschinen, mimographischen Maschinen und Druckerpressen. Das Aufkommen von Elektronik, Computer und optischer Technologie erhöhte die Fähigkeit der Verfasser sogar weiter. Heutzutage kann ein durchschnittlicher Verfasser deutlich mehr Text in einer gegebenen Zeiteinheit erzeugen, als irgendein Verfasser unter Verwendung von handschriftlichen Verfahren der Vergangenheit erzeugen konnte.

[0011] Dieser schnelle Anstieg der Textmenge, gekoppelt mit den enormen technologischen Fortschritten hat bewirkt, daß dem Thema der Übersetzung von Text von seiner Quellensprache in eine Zielsprache (in Zielsprachen) beträchtliche Aufmerksamkeit gezollt wird. Beträchtliche Forschung wurde an Universitäten sowie in privaten und staatlichen Laboratorien betrieben, welche dem Versuch gewidmet war, herauszufinden, wie Übersetzungen ohne das Eingreifen eines menschlichen Übersetzers bewerkstelligt werden können.

[0012] Computerbasierende Systeme wurden entworfen, die die Durchführung einer Maschinenübersetzung (MT = machine translation) versuchten. Solche Computersysteme sind dazu programmiert, das automatische Übersetzen eines Quellentextes als eine Eingabe in einen Zieltext als eine Ausgabe zu versuchen. Jedoch haben Forscher entdeckt, daß solche Computersysteme unmöglich für die Auslegung von automatischer Maschinenübersetzung sind, und zwar unter Verwendung der gegenwärtigen Technologie und des theoretischen Verständnisses. Heutzutage existiert kein System, daß die Maschinenübersetzung einer natürlichen Quellensprache in eine natürliche Zielsprache durchführen kann ohne eine gewisse Art der Editierung durch sachkundige Editoren/Übersetzer. Ein Verfahren wird in der Folge diskutiert.

[0013] In einem Prozeß, der Voreditierung genannt wird, wird ein Quellentext anfänglich durch einen Quelleneditor durchgesehen. Die Aufgabe des Quelleneditors ist es, Veränderungen am Quellentext durchzuführen, um ihn so in Anpassung mit dem zu bringen, was als der Optimalzustand für die Übersetzung durch das Maschinenübersetzungssystem bekannt ist. Diese Anpassung bzw. Übereinstimmung wird vom Quelleneditor über Versuch und Irrtum gelernt.

[0014] Dieser Voreditierprozeß bzw. Vorredigierprozeß, der soeben beschrieben wurde, kann durch Iterationen mittels zusätzlicher Quelleneditoren bzw. Quellenredakteure mit steigender Kompetenz hindurchgehen. Der so vorbereitete Quellentext wird für die Verarbeitung durch das Maschinenübersetzungssystem zugelassen. Die Ausgabe ist ein Zielsprachtext, der, abhängig von den Zwecken der Übersetzung oder den Qualitätsanforderungen des Benutzers, nacheditiert bzw. nachredigiert werden kann oder nicht.

[0015] Wenn die erforderte Übersetzungsqualität vergleichbar sein soll zu geübter menschlicher Übersetzung, muß die Ausgabe der Maschinenübersetzung sehr wahrscheinlich nacheditiert durch einen kompetenten Übersetzer werden. Dies beruht auf der Komplexität der menschlichen Sprache und den vergleichsweise bescheidenen Fähigkeiten der Maschinenübersetzungssysteme, die mit der derzeitigen Technologie gebaut werden können, und zwar innerhalb von natürlichen Einschränkungen bezüglich der Zeit und der Ressourcen und mit einer vernünftigen Erwartung die Anforderungen bezüglich der Kosteneffektivität zu erreichen. Die meisten der bescheidenen Systeme, die gebaut werden, erfordern tatsächlich eine Nacheditiertätigkeit, die dazu dient, sich durch irgendwelche Maßnahmen an die Qualitätspegel der rein menschlichen Übersetzung anzunähern.

[0016] Ein solches System ist das KBMT-89, das vom Center for Machine Translation, Carnegie Mellon University, entworfen wurde, welches Englisch in Japanisch und Japanisch in Englisch übersetzt. Es arbeitet mit einem wissensbasierenden Domänenmodell, das eine interaktive Zweideutigkeitsentfernung unterstützt (d. h. Redigieren des Dokuments, um es unzweideutig zu machen). Jedoch wird diese interaktive Zweideutigkeitsentfernung typischer Weise nicht interaktiv mit dem Verfasser durchgeführt. Sobald das System einen zweideutigen Satz findet, aus dem es nicht die Zweideutigkeit entfernen kann, muß es mit der Verarbeitung stoppen und die Zweideutigkeit durch Befragung eines Verfassers/Übersetzers mittels einer Serie von Auswahlfragen bzw. Multiple-Choice-Fragen auflösen. Zusätzlich, da das KBMT-89 nicht eine gut definierte gesteuerte bzw. kontrollierte Eingabesprache verwendet, erzeugt die sogenannte übersetzerunterstützte interaktive Zweideutigkeitsentfernung einen Text, der eine Nacheditierung erfordert.

[0017] Hinsichtlich des oben genannte wäre es vorteilhaft, ein Übersetzungssystem zur Hand zu haben, das sowohl das Vor- als auch das Nacheditieren eliminiert.

[0018] Im IBM Technical disclosure bulletin, Mai 1986, Band 28, Nr. 12, Seiten 5284–5286 wird ein interaktives verfasserunterstütztes Werkzeug beschrieben, das dem Verfasser erlaubt, in einer natürlichen Sprache geschriebene Dokumente besser übersetzbar zu machen.

[0019] Gemäß der vorliegenden Erfindung wird ein computerbasierendes System für einsprachige Dokumentenentwicklung vorgesehen, das folgendes aufweist:
Einen Texteditor angepaßt zur interaktiven Aufnahme eines Eingabetextes eines Verfassers in einer Quellensprache; und
einen Spracheditor, der eine Erweiterung des Texteditors ist, welcher interaktiv lexikalische Einschränkungen und grammatikalische Einschränkungen auf einen durch den Verfasser verwendeten Teilsatz der natürlichen Sprache erzwingt, um den Eingabetext zu erzeugen, wobei der Verfasser interaktiv bei der Erzwingung der lexikalischen Einschränkungen und der grammatikalischen Einschränkungen auf den Eingabetext unterstützt wird, um so einen unzweideutig eingeschränkten Text zu erzeugen;
ein Maschinenübersetzungssystem, das ansprechend auf den Spracheditor ist, welches zur Übersetzung des unzweideutigen, eingeschränkten Textes in eine Fremdsprache konfiguriert ist; weiter gekennzeichnet durch
einen Domänenmodell, das mit dem Spracheditor kommuniziert, wobei das Domänenmodell ein vorbestimmtes Domänenwissen und ein linguistisches, semantisches Wissen über die lexikalischen Einheiten und deren Kombinationen vorsieht, um so den Spracheditor bei der Erzwingung der lexikalischen und grammatikalischen Einschränkungen zu unterstützen, wobei das Domänenmodell ein dreiteiliges Domänenmodell ist, wobei das dreiteilige Domänenmodell folgendes aufweist:
Einen Kern, der lexikalische Information enthält, die durch den Spracheditor und das Maschinenübersetzungssystem angefordert wird, wobei die lexikalische Information lexikalische Eintragungen bzw. Punkte umfaßt, und zwar innerhalb des Teilsatzes der natürlichen Sprache, zusammen mit assoziierten semantischen Konzepten, Wortklassen und morphologischer Information,
ein Spracheditordomänenmodell, das Information enthält, die nur durch den Spracheditor angefordert wird, wobei die Information zumindest einen Teilsatz der natürlichen Sprache von Synonymen für Eintragungen bzw. Punkte, die nicht innerhalb des Teilsatzes der natürlichen Sprache sind, ein Wörterbuch für Definitionen der lexikalischen Punkte und einen Satz von Beispielen zur Verwendung der lexikalischen Punkte umfaßt, und
ein Maschinenübersetzungsdomänenmodell, das Information enthält, die nur durch das Maschinenübersetzungsdomänenmodell angefordert wird, und zwar einschließlich einer Hierarchie von Konzepten, die für die unzweideutige Abbildung und semantische Verifizierung der Übersetzung verwendet wird.

Kurze Beschreibung der Zeichnungen

[0020] **Fig.** 1(a) und 1(b) sind Blockdiagramme der oberen Architekturebene der vorliegenden Erfindung.

[0021] **Fig.** 2 ist ein Fließdiagramm der hohen Ebene des Betriebs der vorliegenden Erfindung.

[0022] **Fig.** 3 ist der Informationsfluß in der hohen Ebene und ein architektonisches Blockdiagramm des MT **120**.

[0023] **Fig.** 4 zeigt ein Beispiel eines Informationselements.

[0024] **Fig.** 5 ist ein Blockdiagramm des Domänenmodells **500**.

[0025] **Fig.** 6 ist ein Hochebenenfließdiagramm bezüglich des Betriebs des Spracheditors **130**.

[0026] **Fig.** 7 ist ein Fließdiagramm, daß den Betrieb des Vokabelüberprüfers **610** darstellt.

[0027] **Fig.** 8 ist ein Hochebenenfließdiagramm des Zweideutigkeitsentfernungsblocks **630**.

[0028] **Fig.** 9 ist der Informationsfluß und ein architektonisches Blockdiagramm des MT **120**.

Detaillierte Beschreibung der vorliegenden Erfindung

1. Integrierter Systemüberblick

[0029] Das computerbasierende System der vorliegenden Erfindung sieht die funktionale Integrierung von folgendem vor:

1) Eine Verfasserumgebung für die Dokumentenentwicklung, und

2) Ein Modul für eine genaue Maschinenübersetzung in mehrere Sprache ohne Vor- oder Nacheditierung.

Unter Verwendung dieser Technologie bei der Erzeugung von mehrsprachiger Dokumentation wird dem Benutzer eine konsistente genaue, zeitige, kosteneffiziente Übersetzung zugesichert, ob nun in kleinen oder großen Mengen, und zwar mit annähernd simultaner Ausgabe von Information in sowohl der Quellensprache als auch den Sprachen, auf die die Übersetzung gerichtet ist.

[0030] Die Entscheidung, die Quellensprachenverfassungsfunktion mit der Übersetzungsfunktion zu verbinden, basiert auf zwei Prinzipien:

1) In einer multinational, mehrsprachigen Geschäftsumgebung wird Information nicht als vollständig entwickelt angesehen, bis sie in den verschiedenen Sprachen der Benutzer lieferbar ist.

2) Kombinierung der Verfassung- und Übersetzungsprozesse innerhalb eines einheitlichen Rahmenwerks führt zu Effizienzgewinnen, die auf andere Weise nicht erreicht werden können.

[0031] **Fig.** 1(a) zeigt ein Blockdiagramm einer hohen Ebene des integrierten Verfassung- und Übersetzungssystems (IATS = Integrated Authoring and Translation System) **105**. Das IRTS **105** sieht eine spezialisierte Rechnerumgebung vor, die der Unterstützung einer Organisation bei der Verfassung von Dokumentation in einer Sprache und der Übersetzung in verschiedene andere gewidmet ist. Diese zwei verschiedenen Funktionen werden durch eine integrierte Gruppe von Programmen wie folgt unterstützt:

1) Verfassen – eine Sub- bzw. Teilgruppe der Programme sieht einen interaktiven, computerisierten Texteditor (TE) **140** vor, der die Verfasser in die Lage versetzt, ihren einsprachigen Text innerhalb der lexikalischen und grammatikalischen Einschränkungen eines domänengebundenen Teilsatzes einer natürlichen Sprache zu erzeugen, wobei der Teilsatz als die eingeschränkte Quellensprache (CSL = Constrained Source Language) bezeichnet wird. Zusätzlich versetzt der TE **140** die Verfasser bzw. Autoren in die Lage, ferner den Text für eine Übersetzung vorzubereiten, indem er sie durch den Prozeß der Textzweideutigkeitsentfernung führt, welche den Text übersetzbar ohne eine Voreditierung macht;

2) Übersetzung – eine weitere Sub- bzw. Teilgruppe der Programme sieht eine Maschinenübersetzungsfunktion (MT = Machine Translation) **120** vor, die in der Lage ist, die CSL in so viele Zielsprachen zu übersetzen, wie das Erzeugungsmodul bzw. Generatormodul zu deren Erzeugung programmiert wurde, wobei die resultierende Übersetzung keine Nacheditierung erfordert.

[0032] Für ein System, das die Übersetzung als eine zentrale Komponente aufweist, ist die Integrierung des Verfassens und der Übersetzungsfunktionen der vorliegenden Erfindung innerhalb eines einheitlichen Rahmenwerks der einzige Weg, der bis heute entworfen wurde, der sowohl eine Vor- als auch eine Nacheditierung eliminiert.

[0033] Der Texteditor (TE) **140** ist ein Werkzeugsatz zur Unterstützung der Verfasser und Redakteure bzw. Editoren bei der Erzeugung von Dokumenten in der CSL. Diese Werkzeuge werden die Verfasser dabei unterstützen, das geeignete CSL-Vokabular und die Grammatik beim Schreiben ihrer Dokumente zu verwenden. Der TE **140** kommuniziert mit dem Verfasser **160** (und umgekehrt) direkt.

[0034] Bezugnehmend auf **Fig.** 1(b) ist das IATS **105** in vier Hauptteile zur Durchführung der Verfassungs- und Übersetzungsfunktionen unterteilt: (1) eine eingeschränkte Quellensprache (CSL) **133**, (2) ein Texteditor (TE) **140**, (3) ein MT **120** und (4) ein Domänenmodell (DM) **137**. Der Texteditor **140** umfaßt einen Spracheditor **130** und einen Graphikeditor **150**. Zusätzlich ist ebenso ein Dateimanagementsystem (FMS = File Management System) **110** zur Steuerung aller Prozesse vorgesehen.

[0035] Die CSL **133** ist ein Teilsatz der Quellensprache, dessen Grammatik und Vokabular den Bereich bzw.

die Domäne der Dokumentation des Verfasser abdeckt, welche übersetzt werden soll. Die CSL **133** ist mittels Spezifikationen des Vokabulars und grammatikalischer Konstruktionen definiert, die zugelassen sind, so daß der Übersetzungsprozeß ohne die Hilfe einer Vor- und Nacheditierung möglich ist.

[0036] Der TE **140** ist ein Satz von Werkzeugen zur Unterstützung der Verfasser und Redakteure beim Erzeugen von Dokumenten in der CSL. Diese Werkzeuge werden die Verfasser bei der Verwendung des geeigneten CSL-Vokabulars und der Grammatik zum Schreiben ihrer Dokumente unterstützen. Der LE **130** (LE = Language Editor = Spracheditor) kommuniziert mit dem Verfasser **160** (und umgekehrt) über den Texteditor **140**. Der Verfasser verfügt über eine bidirektionale Kommunikation über Leitung **162** mit dem Texteditor **140**. Der LE **130** informiert den Verfasser **160**, ob Worte und Phrasen bzw. Sätze, die verwendet werden, innerhalb der CSL sind. Der LE **130** ist in der Lage, Synonyme in CSL für Worte vorzuschlagen, die relevant für den Informationsbereich sind, der dieses Dokument umfaßt, jedoch nicht in CSL. Zusätzlich unterrichtet der LE **130** den Verfasser **160**, ob oder ob nicht ein Textstück die CSL-grammatikalischen Einschränkungen erfüllt. Er versieht einen Verfasser ebenso mit der Unterstützung bei der Zweideutigkeitsentfernung in Sätzen, die syntaktisch korrekt sein können jedoch semantisch zweideutig sind.

[0037] Der MT **120** ist in zwei Teile unterteilt: Einen MT-Analysierer **127** und einen MT-Generator bzw. -Erzeuger **123**. Der MT-Analysierer **127** dient zwei Zwecken: Er analysiert ein Dokument, um sicherzustellen, daß das Dokument unzweideutig der CSL angepaßt ist, und erzeugt einen Zwischensprachentext (interlingua text). Der analysierte, CSL-gebilligte bzw. -genehmigte Text wird dann in eine ausgewählte Fremd-(Ziel)-Sprache **180** übersetzt. Der MT **120** verwendet eine zwischensprachenbasierende Übersetzungsherangehensweise. Anstatt der direkten Übersetzung eines Dokuments in eine andere Fremdsprache wandelt der MT-Generator **123** das Dokument in eine sprachunabhängige, computerlesbare Form um, die Interlingua bzw. Zwischensprache genannt wird, und erzeugt dann Übersetzungen aus dem Interlinguatext. Als ein Ergebnis werden übersetzte Dokumente keine Nacheditierung erforderlich machen. Eine Version des MT **120** wird für jede Sprache erzeugt und wird im wesentlichen aus einem Satz aus Wissensquellen bestehen, die so entworfen sind, daß sie die Übersetzung des Interlingua- bzw. Zwischensprachentextes in einen Fremdsprachentext führen bzw. leiten. Insbesondere muß für jede neue Zielsprache ein neuer MT-Generator **123** individuell entwickelt werden.

[0038] Bei völliger Funktionsfähigkeit muß der LE **130** manchmal den Verfasser **160** auffordern, aus alternativen Interpretationen für bestimmte Sätze auszuwählen, die die CSL-grammatikalischen Einschränkungen erfüllen, jedoch deren Bedeutung unklar ist. Dieser Prozeß ist als Zweideutigkeitsentfernung bekannt. Nachdem der LE **130** bestimmt hat, daß ein besonderer Teil des Textes nur CSL-Vokabular verwendet und alle CSL-grammatikalischen Einschränkungen erfüllt, dann wird der Text als CSL-genehmigt gekennzeichnet, und zwar bis auf weiteres bzw. schwebend für diese Zweideutigkeitsentfernung. Wie weiter unten erklärt, wird eine Zweideutigkeitsentfernung keine Veränderungen bezüglich der Aspekte des Textes, die durch den Verfasser sichtbar sind, erforderlich machen. Nachdem der Text von Zweideutigkeiten befreit wurde, wird er bereit für die Übersetzung in die Zielsprache **180** sein.

[0039] In der Praxis ist der LE **130** als eine Erweiterung des Texteditors **140** aufgebaut, der die grundlegenden Wortverarbeitungsfunktionalitäten vorsieht, die durch Verfasser und Redakteure zum Erzeugen von Texten und Tabellen angefordert sind. Der Graphikeditor **150** wird zur Erzeugung von Graphiken verwendet. Der Graphikeditor **150** sieht ein Mittel zum Zugriff für Textkennzeichnungen an Graphiken über den Texteditor **140** vor, so daß diese Textkennzeichnungen ebenso CSL-genehmigt werden können.

[0040] Der LE **130** (über Texteditor **140**) kommuniziert mit dem MT-Analysierer **127** und durch diesen mit dem DM **137** während der Zweideutigkeitsentfernung über bidirektionale Einsteckleitungen (socket-to-socket lines). Im bevorzugten Ausführungsbeispiel der vorliegenden Erfindung ist das DM eine der Wissensbasen, die den MT-Analysierer **120** versorgen. Das DM **137** ist eine symbolische Repräsentation des deklarativen bzw. Erklärungswissens über das CSL-Vokabular, das durch den MT-Analysierer **127** und den LE **130** verwendet wird.

[0041] **Fig.** 2 zeigt ein Hochebenenfließdiagramm des Betriebs des IATS **105**. Der MT **120**, der LE **130**, der Texteditor **140** und der Graphikeditor **150** werden alle durch das FMS **110** gesteuert. Steuerleitungen **111**–**113** sehen die notwendige Steuerinformation für den richtigen Betrieb des IATS **105** vor.

[0042] Anfänglich wird der Verfasser **160** das FMS **110** verwenden, um ein zu editierendes bzw. zu redigierendes Dokument auszuwählen, und das FMS **110** wird den Texteditor **140** starten, und zwar die Datei für das spezifizierte Dokument anzeigend. Über den Texteditor **140** gibt der Verfasser einen Text ein, der ein uneingeschränkter und zweideutiger Text sein kann, und zwar in das IATS **105**, wie in den Blöcken **160** und **220** gezeigt. Der Verfasser bzw. Autor **160** wird Standardeditorbefehle verwenden, um das Dokument zu erzeugen

und zu modifizieren, und zwar bis es bereit ist, um für eine CSL-Konformität bzw. -Anpassung überprüft zu werden. Es sei bemerkt, daß erwartet wird, daß Verfasser vorwiegend Text eingeben werden, der im wesentlichen hinsichtlich der CSL-Einschränkungen vorbereitet ist. Der Text wird dann durch den Verfassern ansprechend auf eine Systemrückkopplung modifiziert werden, und zwar basierend auf Verletzungen von vorbestimmten lexikalischen und grammatikalischen Einschränkungen, um konform mit der CSL zu sein. Dies ist natürlich viel effizienter, als anfänglich völlig uneingeschränkten Text einzugeben. Jedoch wird das System richtig funktionieren bzw. arbeiten, sogar wenn völlig uneingeschränkter Text vom Beginn an eingegeben wird. Die Kommunikation des Verfassers mit dem LE **130** besteht aus Befehlen durch Klicken der Maus oder Tastendrücken. Jedoch sollte bemerkt sein, daß andere Formen der Eingabe benutzt werden können, wie beispielsweise die Verwendung eines Griffels, der Stimme, usw., jedoch nicht auf deren Verwendung eingeschränkt, ohne den Umfang oder die Funktion der vorliegenden Erfindung zu verändern. Ein Beispiel einer Eingabe ist ein Befehl eine CSL-Überprüfung durchzuführen oder die Definition und ein Verwendungsbeispiel für ein gegebenes Wort oder eine Phrase bzw. einen Ausdruck zu finden.

[0043] Der CSL-Text, der restliche Zweideutigkeiten oder stilistische Probleme enthalten kann, wird auf ein Konformität mit der CSL analysiert und bezüglich der Einhaltung von grammatikalischen Regeln überprüft, die in der Wissensbasis enthalten sind, wie in Block **230** gezeigt. Der Autor wird mit einer Rückkopplung versehen, um jeglichen Fehler über die Rückkopplungsleitung **215** zu korrigieren. Im Speziellen sieht der LE **130** Information bezüglich von Nicht-CSL-Worten und -Ausdrücken bzw. -Phrasen und -Sätzen für den Verfasser **160** vor. Zuletzt wird der Text auf irgendwelche zweideutigen Sätze überprüft. Der LE fordert den Verfasser auf, eine geeignete Interpretation eines Satzinhaltes bzw. einer Satzbedeutung auszuwählen. Dieser Prozeß wird wiederholt bis der Text vollständig von Zweideutigkeiten befreit ist.

[0044] Sobald der Autor alle notwendigen Korrekturen am Text angebracht hat und die Analysephase **230** abgeschlossen ist, wird der zweideutigkeitsbefreite/eingeschränkte Text **240** an den MT-Analysierer und -Übersetzer **250** weitergegeben. Der Interpretierer bzw. die Übersetzungsvorrichtung befindet sich im MT-Analysierer **127** zusammen mit dem syntaktischen Teil des Analysierers und übersetzt den zweideutigkeitsbefreiten/eingeschränkten Text **240** in Interlingua bzw. die Zwischensprache **260**. Die Zwischensprache **260** wird wiederum durch den Generatorblock **270** in den Zieltext **280** übersetzt. Wie in **Fig.** 3 gezeigt, liegt der Zwischensprachentext **260** in einer Form vor, die in mehrere Zielsprachen **306**–**310** übersetzt werden kann.

[0045] Dadurch, daß der Verfasser aufgefordert wird und in die Lage versetzt wird, Dokumente zu erzeugen, die konform mit spezifischem Vokabular und grammatikalischen Einschränkungen sind, ist es möglich, die genaue Übersetzung des Textes mit eingeschränkter Sprache in Fremdsprachen durchzuführen, ohne daß eine Nacheditierung erforderlich ist. Eine Nacheditierung ist deshalb nicht erforderlich, weil der LE-Vokabularüberprüfungsblock **217** und der Analyseblock **230** den Verfasser veranlaßt haben, alle möglichen zweideutigen Sätze und alle nicht übersetzbaren Worte aus dem Dokument vor der Übersetzung zu modifizieren und/oder von Zweideutigkeiten zu befreien.

II. Detaillierte Beschreibung der Funktionsblöcke

[0046] In einem bevorzugten Ausführungsbeispiel wird jeder Verfasser die Nutzung einer einzigen DEC-Station mit 32 Mega RAM, einem 400-Megabytediskettenlaufwerk und einem 19-Zollfarbmonitor haben. Jede Workstation bzw. Arbeitsstation wird für einen Austausch von zumindest 100 Meg von seiner lokalen Platte bzw. Festplatte konfiguriert sein. Zusätzlich zu den Workstations der Verfasser werden DEC-Server als File- bzw. Datenserver verwendet werden, und zwar jeweils einer für jeweils zwei Verfassergruppen, und zwar für insgesamt nicht mehr als 45 Benutzer pro Dateiserver. Des weiteren befinden sich die Verfasserworkstations in einem Ethernet lokalen Netzwerk. Das System verwendet das Unixbetriebssystem (eine Abwandlung gemäß der Berkeley Standardverteilung (BSD = Berkeley Standard Distribution) ist gegenüber einer Abwandlung gemäß System V (SYSV) bevorzugt. Ein Compiler für C-Programmiersprache und OSF/Motiv-Bibliotheken sind verfügbar. Der LE wird innerhalb eines Motiv-Fenstermanagers laufen gelassen. Es sei bemerkt, daß die vorliegende Erfindung nicht auf die zuvor genannte Hardware- und Softwareumgebungen eingeschränkt ist, und andere Umgebungen werden für die vorliegende Erfindung in Betracht gezogen.

A. Texteditor

[0047] Das bevorzugte Ausführungsbeispiel der vorliegenden Erfindung sieht einen Texteditor **140** vor, der dem Verfasser erlaubt, Information einzugeben, die letztendlich analysiert und schließlich in eine Fremdsprache übersetzt wird. Jegliche kommerziell erhältliche Wortverarbeitungssoftware kann mit der vorliegenden Erfindung verwendet werden. Ein bevorzugtes Ausführungsbeispiel verwendet einen SGML-Texteditor **140**, der

von ArborText geliefert wird (ArborText Inc., 535 West William St., Ann Arbor, MI 48103). Der SGML-Texteditor **140** sieht eine Basiswortverarbeitungsfunktionalität vor, die von den Verfassern und Redakteuren angefordert ist, und er wird mit Software von InterCap (von Annapolis, Maryland) zur Erzeugung von Graphiken verwendet.

[0048] Die vorliegende Erfindung verwendet einen SGML-Texteditor **140**, da er Text unter Verwendung von Kennzeichnungen in Standard Generalized Markup Language (SGML = standardisierte verallgemeinerte Markupsprache) erzeugt. SGML ist eine internationale standardisierte Markup- bzw. Auszeichnungssprache zur Beschreibung der Struktur von elektronischen Dokumenten. Sie ist konstruiert, um die Erfordernisse für einen weiten Bereich der Dokumentenverarbeitung und Austauschaufgaben zu erfüllen. SGML-Kennzeichnungen ermöglichen es, daß Dokumente bezüglich ihres Inhalts (Text, Bilder, usw.) und ihrer logischen Struktur (Kapitel, Absätze, Figuren, Tabellen, usw.) beschrieben werden. Im Falle von größeren, komplexeren, elektronischen Dokumenten macht sie es ebenso möglich, daß die physikalische Organisation eines Dokuments in Dateien beschrieben wird. SGML ist konstruiert, um zu ermöglichen, daß Dokumente irgendeines Typs, einfach oder komplex, kurz oder lang, in einer Weise beschrieben werden, die unabhängig sowohl vom System als auch der Anwendung ist. Diese Unabhängigkeit ermöglicht einen Dokumentenaustausch zwischen verschiedenen Systemen für verschiedene Anwendungen ohne eine Fehlinterpretation oder den Verlust von Daten. SGML ist eine Markup- bzw. Auszeichnungssprache, d. h. eine Sprache zum "Auszeichnen" bzw. "Kennzeichnen" oder Kommentieren von Text mittels oder durch die Verwendung von kodierter Information, die zur herkömmlichen Textinformation hinzugefügt wird, die durch ein gegebenes Textstück vermittelt wird. In den meisten Fällen nimmt sie die Form von Abfolgen bzw. Sequenzen von Zeichen an verschiedenen Punkten durch ein elektronisches Dokument hindurch an. Jede Sequenz ist unterscheidbar vom sie umgebenden Text durch spezielle Zeichen, die sie beginnen und sie beenden. Die Software kann verifizieren, daß das korrekte Markup bzw. die korrekte Auszeichnung in den Text eingefügt wurde, und zwar durch Prüfung der SGML-Kennzeichnungen bei Anfrage. Das Markup ist verallgemeinert dahingehend, daß es nicht spezifisch für irgendein besonderes System oder eine Aufgabe ist. Für eine tiefere Diskussion von SGML-Kennzeichnungen siehe International Standard (ISO) 8879, Information processing – Text and office systems – Standard Generalized Markup Language (SGML), Ref. No. ISO 8879-1986(E).

[0049] Die folgenden Fähigkeiten sind möglich aufgrund der Verwendung von SGML-Kennzeichnungen:

(1) Unterteilung des Dokuments in Fragmente oder übersetzbare Einheiten. Die Texteditor-140-Software verwendet sowohl Zeichensetzung als auch SGML-Kennzeichnungen, um übersetzbare Einheiten im Quelleneingabetext zu erkennen (d. h. ein SGML-Kennzeichen ist notwendig, um Abschnittstitel zu identifizieren);

(2) Abschirmen (Isolieren) von Einheiten, die nicht übersetzt werden. Obwohl das System auf der Prämisse bzw. Vorgabe basiert, daß alle Worte und Sätze der eingeschränkten Sprache angehören, die nicht im voraus vorhergesagt werden können (beispielsweise Namen und Adressen), oder Vokabularklassen, die nicht (bereitwillig) ausreichend bzw. erschöpfend spezifiziert werden können (beispielsweise Teilenummern, Fehlermeldungen von Maschinen). SGML-Kennzeichnungen können um diese Eintragungen bzw. Punkte herum gegeben werden, um dem System anzuzeigen, daß sie aus der Überprüfung herausgenommen werden;

(3) Identifizierung des Inhalts (d. h. Teilenummer) wie unter (2) diskutiert;

(4) Erlauben der Übersetzung von Teilsätzen (d. h. mit Kennzeichen versehene Eintragungen bzw. Punkte (bulleted items));

(5) Unterstützung der Übersetzung von Tabellen (jeweils eine Zelle) durch Identifizierung der Textstruktur. Dieses Merkmal ist ähnlich zu dem unter (1) beschriebenen;

(6) Unterstützung des Prozesses zur Satzteilbestimmung (parsing) (weiter unten beschrieben) über (2), (3), (4), (5);

(7) Unterstützung der Zweideutigkeitsentfernung durch Vorsehen eines Mittels zum Einfügen von unsichtbaren Kennzeichnungen in den Quellentext, so daß die korrekte Interpretation eines zweideutigen Satzes angezeigt wird;

(8) Unterstützung der Übersetzung von Währungen und mathematischen Einheiten über Identifizierung von spezifischen Texttypen, die eine spezielle Behandlung erfordern.

(9) Vorsehen von Mitteln zur Kennzeichnung eines Textteiles als übersetzbar. In anderen Worten wird zertifiziert bzw. bestätigt, daß ein Textteil durch den Prozeß hindurch geschritten ist, der weiter unten umrissen wird, und daß der Text ein unzweideutiger, eingeschränkter Text ist, der ohne eine Nacheditierung übersetzt werden kann.

[0050] In der Vergangenheit haben Verfasser (mittels des Texteditors **140**) elektronische Dokumente (nur Text – keine Graphiken) erzeugt, der ein vollständiges "Buch" repräsentierten. Dies impliziert, daß die gesamte Arbeit durch einen Schreiber getan wurde, und daß die erzeugte Information nicht einfach wieder verwendet wird.

Die vorliegende Erfindung kompiliert jedoch (oder erzeugt) Bücher (Betriebsanleitungen, Dokumente) aus einem Satz von kleineren Stücken oder Informationselementen, was impliziert, daß die Arbeit von mehreren Schreibern getan werden kann. Das Ergebnis dieser Erfindung ist eine verbesserte Wiederverwertbarkeit bzw. Wiederverwendbarkeit. Ein Informationselement wird als das kleinste alleinstehende Stück einer Dienstinformation über eine spezialisierte Domäne bzw. einen spezialisierten Bereich definiert. Es sei bemerkt, daß jedoch, obwohl ein bevorzugtes Ausführungsbeispiel Informationselemente verwendet, die vorliegende Erfindung richtige, unzweideutig übersetzte Dokumente ohne die Verwendung von Informationselementen erzeugen kann.

[0051] **Fig.** 4 zeigt ein Beispiel eines Informationselements **410**, das eine "einzigartige" Überschrift **415**, einen "einzigartigen" Textblock **420**, eine "geteilte" bzw. "gemeinsame" Graphik **430**, eine "geteilte" Tabelle **435** und einen "geteilten" Textblock **425** umfaßt. "Einzigartige" Information ist jene Information, die nur auf das Informationselement anwendbar ist, in welchem sie enthalten ist. Dies impliziert, daß die "einzigartige" Information als Teil des Informationselements **450** geführt wird. Ein "geteiltes" bzw. "gemeinsames" Objekt (eine Graphik, Tabelle oder ein Textblock) ist Information, auf die "Bezug genommen" wird im Informationselement. Der Inhalt von "geteilten" Objekten wird im Verfasserwerkzeug dargestellt, jedoch wird auf ihn nur "gezeigt" im aufgeführten bzw. eingereichten Informationselement **450**. "Geteilte" Objekte unterscheiden sich von Informationselementen dahingehend, daß sie nicht alleinstehend sind (d. h. sie können nicht genug Information selbst vermitteln, um substantielle Information beizutragen). Jedes "geteilte" Objekt ist selbst eine getrennte Datei bzw. Ablage, wie in Block **450** gezeigt.

[0052] Informationselemente werden durch Kombination von "einzigartigen" Informationsblöcken (Text und/oder Tabellen) mit einem oder mehreren "geteilten" Objekten gebildet. Es sei bemerkt, daß eine "einzigartige" Überschrift **415** und ein "einzigartiger" Text **420** mit der "geteilten" Graphik **430** der "geteilten" Tabelle **435** und dem "geteilten" Text **425** kombiniert werden. Ein Satz von einem oder mehreren Informationselementen macht ein vollständiges Dokument (Buch).

[0053] "Geteilte" Objekte werden in "geteilten" Bibliotheken gespeichert. Bibliothektypen umfassen "geteilte" Graphikbibliotheken **460a**, "geteilte" Tabellenbibliotheken **460b**, "geteilte" Textbibliotheken **460c**, "geteilte" Audiobibliotheken **460d** und "geteilte" Videobibliotheken **460e**. Ein geteiltes Objekt wird nur einmal gespeichert. Wenn es in einzelnen bzw. individuellen Informationselementen verwendet wird, werden nur "Zeiger" zum originalen geteilten Objekt in die Datei bzw. die Ablage **450** mit geteilter Information plaziert. Dies minimiert den Betrag des Festplattenplatzes bzw. Diskettenplatzes, der erforderlich ist. Wenn das originale Objekt verändert wird, werden all jene Informationselemente, die auf dieses Objekt "zeigen" automatisch verändert. Ein geteiltes Objekt kann in jeglichen Publikationstyp verwendet werden. Ein "geteiltes Informationselement" ist ein Informationselement, das in mehr als einem Dokument verwendet wird. Beispielsweise werden die selben vier Informationselemente in der Freigabebibliothek **470** verwendet, um Teile der Dokumente **480** und **485** zu erzeugen.

[0054] Alle Kommunikation zwischen dem Verfasser und dem LE **130** wird über ein LE-Benutzerinterface bzw. eine LE-Benutzerschnittstelle (UI) vermittelt, die entweder als eine Erweiterung der Standard-SGML-Editormöglichkeiten implementiert ist, wie beispielsweise Menüoptionen, oder in separaten bzw. getrennten Fenstern implementiert ist. Die UI sieht einen Zugriff auf und eine Steuerung der CSL-Überprüfer und das CSL-Vokabularnachschlagewerk vor und managt dasselbe, und sie ist das vorrangige bzw. Primärwerkzeug, das Benutzer zur Interaktivität mit dem CSL-LE befähigt. Obwohl der Ausdruck "Benutzerinterface" oft in einem allgemeineren Sinn verwendet wird, um sich auf die Schnittstelle bzw. das Interface zu einem gesamten Softwaresystem zu beziehen, ist hier der Ausdruck darauf beschränkt, daß der die Schnittstelle zum CSL-Überprüfer, zur Vokabularnachschlagfähigkeit bzw. -einrichtung und zur Zweideutigkeitsentfernungsfähigkeit bzw. -einrichtung bedeutet.

[0055] Neben anderen Dingen muß die UI eine klare Information vorsehen bezüglich (a) den Aktionen bzw. Tätigkeiten, die der LE aufnimmt, (b) dem Ergebnis dieser Tätigkeiten und (c) jeglicher Folgetätigkeiten. Beispielsweise wann immer eine Tätigkeit, die durch die UI initiiert wurde, mehr als eine sehr kurze Realzeitpause einführt, sollte die UI den Verfasser über mögliche Verzögerungen mittels einer kurzen bzw. knappen Nachricht informieren. Der Verfasser kann LE-Funktionalitäten durch Auswahl einer Option aus einem Pulldown-Menü bzw. Ziehmenü im Texteditor **140** auswählen. Die verfügbaren Optionen erlauben dem Verfasser, eine Rückkopplung von der CSL-Überprüfung (sowohl Vokabular- als auch Grammatiküberprüfung) und von der Vokabularnachschlagung zu initiieren und zu sichten. Der Autor bzw. Verfasser kann anfordern, daß eine Überprüfung auf dem derzeit dargestellten Dokument initiiert wird oder er kann eine Vokabularnachschlagung bezüglich eines gegebenen Worts oder eines Ausdrucks bzw. einer Phrase anfordern. Die UI wird klar jedes Auftre-

ten von Nicht-CSL-Sprache anzeigen, das im Dokument gefunden wurde. Mögliche Wege der Anzeige von Nicht-CSL-Sprache umfassen die Verwendung von Farbe und Veränderungen des Schrifttyps oder der Größe im SGML-Editorfenster. Die UI wird alle bekannte Information bezüglich irgendeines Nicht-CSL-Wortes anzeigen. Beispielsweise wird in geeigneten Fällen die UI eine Nachricht anzeigen, die aussagt, daß das Wort Nicht-CSL ist, jedoch ein CSL-Synonym besitzt, sowie eine Liste dieser Synonyme.

[0056] In Fällen, wo ein Vokabularüberprüfungsbericht eine Liste von Alternativen für Nicht-CSL-Worte umfaßt, auf die sich konzentriert wird (beispielsweise Buchstabieralternativen oder CSL-Synonyme), wird der Verfasser in die Lage versetzt, eine dieser Alternativen auszuwählen und anzufordern, daß sie automatisch in das Dokument ersetzt wird. In einigen Fällen kann es sein, daß der Verfasser die ausgewählte Alternative modifizieren muß (beispielsweise Hinzufügung der richtigen Endung), um sicherzustellen, daß sie in der richtigen Form ist.

[0057] Wenn ein Verfasser Vokabularinformation anfordert, stellt die UI Buchstabieralternativen, Synonyme, eine Definition und/oder ein Verwendungsbeispiel für die angezeigte Eintragung bzw. den Punkt dar.

[0058] Der Verfasser kann sich schnell und leicht zwischen der Überprüfungsinformation und der Vokabularnachschlaginformation innerhalb der UI bewegen. Dies ermöglicht dem Verfasser, Informationssuchen (im allgemeinen Synonymnachschlagungen) während des Prozesses der Veränderung der Dokumente durchzuführen, um Nicht-CSL-Sprache zu entfernen. In den meisten Fällen sieht die UI eine automatische Ersetzung von Nicht-CSL-Vokabular gegen CSL-Vokabular vor, und zwar ohne die Notwendigkeit für den Benutzer, daß CSL-Wort zu modifizieren, um sicherzustellen, daß es in der geeigneten bzw. richtigen Form ist. Jedoch gibt es einige Fälle, in denen die Vokabularüberprüfung (weiter unten beschrieben), die keine Satzteilüberprüfung eines Dokuments durchführt, nicht in der Lage ist, die korrekte, zu liefernde Form zu identifizieren. Man betrachte die folgende Unterschrift bzw. Bildunterschrift, in dem Fall, wo das Verb "view" nicht in CSL ist, jedoch das CSL-Synonym "see" besitzt:
Direction of Crankshaft Rotation
(when viewed from flywheel end)

[0059] Der Vokabularüberprüfer wird nicht wissen, ob "saw" oder "seen" als ein Synonym für "viewed" angeboten werden soll. Natürlich wäre in diesem Fall ein vernünftiger Tätigkeitsablauf das Anbieten von beiden Möglichkeiten und dem Verfasser zu erlauben, die geeignete auszuwählen. Da es keine Sicherheit gibt, daß jeder Fall eine Präsentation erlauben wird, die den Verfasser in die Lage versetzt eine direkte Ersetzung anzuordnen, sieht der LE **130** eine Liste von Ersetzungsoptionen in die korrekte Form vor, wo immer möglich. Es kann jedoch Fälle geben, wo der Verfasser es für notwendig finden wird, ein vorgeschlagenes CSL-Wort oder einen CSL-Ausdruck zu redigieren bzw. zu editieren, bevor angefordert wird, daß es in das Dokument gegeben wird.

[0060] Schließlich sieht die LE UI eine Unterstützung für die Zweideutigkeitsentfernung der Bedeutung von Sätzen vor. Sie bewerkstelligt dies durch Vorsehen einer Liste von möglichen alternativen Interpretationen für den Verfasser, ferner erlaubt sie dem Verfasser die Auswahl der richtigen Interpretation und kennzeichnet dann den Satz, um so die Auswahl des Verfassers anzuzeigen.

B. Dateimanagementsystem

[0061] Das Dateimanagementsystem (FMS = File Management System) **110** dient als die Schnittstelle für den Verfasser zur IE-Freigabebibliothek bzw. - Herausgabebibliothek **470** und zum SGML-Texteditor **140**. Typischer Weise werden Verfasser einen IE zum Editieren auswählen, und zwar durch Anzeigen der Datei für diesen IE in der FMS-Schnittstelle. Die FMS **110** wird dann eine SGML-Editorsitzung für diesen IE initiieren und verwalten. Vollendete Dokumente werden weitergeleitet zu einem menschlichen Editor bzw. Redakteur oder einem Informationsintegrator über FMS-gesteuerte Einrichtungen.

C. Eingeschränkte Quellensprache (CSL)

[0062] Mit der Komplexität von heutiger technischer Dokumentation ist eine hochqualitative Maschinenübersetzung mit natürlicher Sprache von uneingeschränkten Texten praktisch unmöglich. Die Haupthindernisse dahingehend sind von linguistischer Natur. Der Kern- bzw. Schlüsselprozeß bei der Übersetzung eines Quellentextes ist der der Wiedergabe seiner Bedeutung in der Zielsprache. Weil die Bedeutung unter der Oberfläche von Textsignalen liegt, müssen solche offenen bzw. zutageliegenden Signale analysiert werden. Die aus dieser Analyse resultierende Bedeutung wird im Prozeß zur Erzeugung der Signale der Zielsprache verwendet. Eini-

ge der lästigsten bzw. unangenehmsten Übersetzungsprobleme resultieren aus diesen Merkmalen, die der Sprache inhärent sind, welche die Analyse und Erzeugung behindern.

[0063] Einige dieser Merkmale sind:

1. Wörter mit mehr als einer Bedeutung in einem zweideutigen Zusammenhang
Beispiel: Make it with light material.
[Ist das Material "not dark" oder "not heavy"?]
2. Wörter in zweideutigem Auftreten
Beispiel: Das deutsche Wort für "Arbeiterinformation" ist entweder "information for workers" [Arbeiter + Information] oder "formation of female workers" [Arbeiterin + Formation]
3. Wörter, die mehr als eine syntaktische Rolle spielen
"Round" kann ein Hauptwort (N), ein Verb (V) oder ein Adjektiv (A) sein:
(N) Liston was knocked out in the first round.
(V) Round of the figures before tabulating them.
(A) Do not place the cube in a round box.
4. Kombinationen von Wörtern, die jeweils mehr als eine syntaktische Rolle spielen können.
Beispiel: British Left Waffles on Falklands.
[Wenn "Left Waffles" als N + V gelesen wird, dann betrifft die Überschrift die "British Left"]
[Wenn "Left Waffles" als V + N gelesen wird, so betrifft die Überschrift die "British"]
5. Wortkombinationen in zweideutigen Strukturen
Beispiel: Visiting relatives can be boring.
[Geht es um "visiting of relatives" oder um "relatives who visit", was "boring" sein kann]
Beispiel: Lift the head with the lifting eye.
[Ist das "lifting eye" ein Instrument oder ein Merkmal des "head"?]
6. Verwechslung von Pronomenbezug
Beispiel: The monkey ate the banana because it was ...
[Auf was bezieht sich "it" zurück, "the monkey" oder "the banana"?]

[0064] Erzeugungsprobleme addieren sich zu dem Vorhergesagten dazu, was die Gesamtschwierigkeit der Maschinenübersetzung erhöht.

[0065] Die Größe der Übersetzungsprobleme wird bedeutend geschmälert durch eine Reduktion des Bereichs von linguistischen Phänomenen, die die Sprache darstellt. Eine Teilsprache deckt den Bereich von Objekten, Prozessen und Beziehungen innerhalb einer eingeschränkten Domäne bzw. eines Bereichs ab. Auch wenn eine Teilsprache bezüglich ihres Lexikons eingeschränkt sein kann, muß sie nicht notwendigerweise bezüglich der Leistung ihrer Grammatik eingeschränkt sein. Bei kontrollierten bzw. gesteuerten Situationen ist eine Strategie, die auf die Verwendung von Maschinenübersetzungen gerichtet ist, die, daß sowohl das Lexikon als auch die Grammatik der Teilsprache eingeschränkt sind.

[0066] Einschränkungen bezüglich des Lexikons limitieren seine Größe durch Vermeidung von Synonymen und steuern lexikalische Zweideutigkeiten durch Spezialisierung der lexikalischen Einheiten für das Ausdrücken von, soweit möglich, einer Bedeutung pro Einheit. Es ist einfach, sich vorzustellen, wie diese Einschränkungen Probleme vermeiden können, die unter 1, 2 und 4 beispielhaft dargestellt wurden, wie zuvor erwähnt. Grammatikalische Einschränkungen können einfach Prozesse ausschließen, wie beispielsweise Pronomenverwendung (6 von zuvor), oder sie können es erforderlich machen, daß die gemeinte Bedeutung klarer gemacht wird entweder durch Zufügung oder Wiederholung von ansonsten redundanter Information oder durch Überschreiben. Die folgenden Beispiele setzen die Parameter für die Anwendung dieses Erfordernisses:
Uneingeschränktes, zweideutiges Englisch (welches entweder als A, B1 oder B2, wie weiter unten, interpretiert werden kann):
Clean the connecting rod and main bearings.
Unzweideutige Englischversion A:
Clean the connecting rod bearings and the main bearings.
Unzweideutige Englischversion B1:
Clean the main bearings and the connecting rod.
Unzweideutige Englischversion B2:
Clean the main bearings and the connecting rods.

[0067] Die Anzahl und Typen der lexikalischen und grammatikalischen Einschränkungen können über einen weiten Bereich variieren, und zwar abhängig vom Ziel der Entwicklung der eingeschränkten Teilsprache.

[0068] Hinsichtlich des oben genannten limitiert die vorliegende Erfindung das Verfassen von Dokumenten innerhalb der Grenzen einer eingeschränkten Sprache. Eine eingeschränkte Sprache ist eine Teilsprache einer Quellensprache (beispielsweise Amerikanisches Englisch), die für die Domäne bzw. den Bereich einer besonderen Benutzeranwendung entwickelt wurde. Für eine allgemeine Diskussion von eingeschränkten oder kontrollierten Sprachen siehe Adriaens et al, From COGRAM to ALCOGRAM: Toward a controlled English Grammar Checker, Proc. of Coling-92, Nantes (August 23–28, 1992). Im Kontext der Maschinenübersetzung sind die Ziele der eingeschränkten Sprache wie folgt:

1. Ermöglichung eines konsistenten Verfassens von Quellendokumenten und Ermutigung für klares und direktes Schreiben; und
2. Vorsehen eines Prinzipienrahmenwerks für Quellentexte, welches eine schnelle, genaue und hochqualitative Maschinenübersetzung für Benutzerdokumente erlaubt.

[0069] Der Satz von Regeln, die Verfasser befolgen müssen, um sicherzustellen, daß die Grammatik, von dem was sie schreiben, konform mit CSL ist, werden als CSL-grammatikalische Einschränkungen bezeichnet. Die Computerausführung der CSL-grammatikalischen Einschränkungen, die zur Analyse der CSL-Texte in der MT-Komponente verwendet werden, werden als die CSL-Funktionalgrammatik bezeichnet, und zwar basierend auf den gut bekannten Formalismen, die von Martin Kay entwickelt wurden und später durch R. Kaplan und J. Bresnan modifiziert wurden (siehe Kay, M., "Parsing in Functional Unification Grammar," in D. Dowty, L Karttunen und A. Zwicky (eds.), Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives, Cambridge, Mass.: Cambridge University Press, Seiten 251–278 (1985) und Kaplan R. und J. Bresnan, "Lexical Functional Grammar: A Formal System for Grammatical Representation," in J. Bresnan (ed.), The Mental Representation of Grammatical Relations, Cambridge, Mass.: MIT Press, Seiten 172–281 (1982).

[0070] Im Verbleibenden dieses Dokuments beziehen wir uns häufig auf den Ausdruck, daß ein Wort oder eine Phrase "in CSL" oder "nicht in CSL" sein kann. Weiter unten werden wird die Annahme über den Typ der vokabularen Einschränkungen beschreiben, die durch CSL auferlegt werden, und werden die Verwendung des Ausdrucks "in CSL" klarstellen.

[0071] Dasselbe Wort oder dieselbe Phrase in Englisch kann viele verschiedene Bedeutungen haben: Beispielsweise kann ein allgemeines Wörterbuch die folgenden Definitionen für das Wort "leak" aufführen:

(1) Verb: Das Entkommen von etwas durch einen Bruch oder Riß zu erlauben;
(2) Verb: Information offenbaren, ohne offizielle Genehmigung oder Anweisung; und
(3) Hauptwort: Ein Riß oder eine Öffnung, die Etwas erlaubt, aus einem Behälter oder einer Leitung zu entkommen oder in diese einzudringen.

[0072] Auf jede dieser verschiedenen Bedeutungen wird sich als "Sinn" des Wortes oder der Phrase bezogen. Mehrfache Sinngebungen für ein einzelnes Wort oder eine Phrase können Probleme für ein MT-System hervorrufen, welches nicht das Wissen besitzt, das Menschen zum Verständnis verwenden, welche von mehreren möglichen Sinngebungen in einem gegebenen Satz beabsichtigt ist. Für viele Worte kann das System einige Zweideutigkeiten durch Erkennung der Wortklassen des Wortes, wie es in einem besonderen Satz verwendet wird (Hauptwort, Verb, Adjektiv, usw.), eliminieren. Dies ist möglich, weil jede Definition eines Wortes besonders für die Verwendung dieses Wortes als eine bestimmte Wortklasse ist, wie zuvor unter Bezugnahme auf "leak" angedeutet.

[0073] Um jedoch die Arten der Zweideutigkeit zu vermeiden, die der MT **120** nicht eliminieren kann, strebt die CSL-Spezifikation dahin, einen einzigen Sinn eines Wortes oder einer Phrase für jede Wortklasse zu umfassen. Daher, wenn ein Wort oder eine Phrase "in CSL" ist, kann es in CSL mit zumindest einer seiner möglichen Sinngebungen verwendet werden. Beispielsweise kann einem Verfasser, der in CSL schreibt, die Verwendung von "leak" mit den Sinngebungen (1) und (3) von zuvor erlaubt werden, jedoch nicht in der Sinngebung (2). Die Aussage, daß ein Wort oder eine Phrase "in CSL" ist, bedeutet nicht, daß alle möglichen Verwendungen des Wortes oder der Phrase übersetzt werden können.

[0074] Wenn ein Wort oder eine Phrase in CSL ist, dann sind alle Formen dieses Worts oder der Phrase, die seine CSL-Sinngebung (N) ausdrücken können, ebenso in CSL. Im obigen Beispiel kann der Verfasser nicht nur das Verb "leak" sondern auch die in Beziehung stehenden Verbformen "leaked", "leaking" und "leaks" verwenden. Wenn ein Wort oder eine Phrase mit einer Hauptwortsinngebung Teil von der CSL ist, sind sowohl seine Singular- als auch die Pluralformen verwendbar. Man bemerke jedoch, daß Phrasen bzw. Ausdrücke, die mehr als eine Wortklassenfunktion haben, ungewöhnlich sind. Diese Heuristik ist daher weniger relevant im Fall einer zweideutigen Phrase.

[0075] Ein Vokabular ist die Sammlung von Worten und Phrasen, die in einer besonderen Sprache oder Teilsprache verwendet werden. Auf eine eingeschränkte Domäne bzw. einen eingeschränkten Bereich wird sich mittels eines eingeschränkten Vokabulars bezogen, welches zur Kommunikation oder zum Ausdruck von Information über einen eingeschränkten Raum der Erfahrung verwendet wird. Ein Beispiel einer eingeschränkten Domäne könnte beispielsweise Landwirtschaft sein, wobei das eingeschränkte Vokabular Begriffe umfassen würde, die landwirtschaftliche Ausrüstungen und Aktivitäten betreffen. Die MT-Komponente wird auf mehr als einer Vokabularart operieren. Die Wörter und Phrasen für eine Maschinenübersetzung werden im MT-Lexikon gespeichert. Das Vokabular kann in verschiedene Klassen unterteilt werden: (1) funktionale Eintragungen; (2) Eintragungen mit allgemeinem Inhalt; und (3) technische Nomenklatur.

[0076] Funktionale Eintragungen in Englisch sind einzelne Worte und Wortkombinationen, die primär zur Verknüpfung von Ideen in einem Satz dienen. Sie sind für nahezu jeglichen Typ von geschriebener Kombination in Englisch erforderlich. Diese Klasse umfaßt Präpositionen (to, from, with, in front of, usw.), Konjunktionen (and, but, or, if, when, because, since, while, usw.), Bestimmungswörter (the, a, your, most of), Pronomen (it, something, anybody, usw.) einige Adverbien (no, never, always, not, slowly, usw.), und Hilfsverben (should, may, ought, must, usw.).

[0077] Wörter mit allgemeinem Inhalt werden im großen Umfang zur Beschreibung der uns umgebenden Welt verwendet; ihre Hauptverwendung ist die Reflektion der gewöhnlichen und gemeinsamen menschlichen Erfahrung. Typischerweise sind Dokumente auf einen sehr spezialisierten Teil der menschlichen Erfahrung fokussiert (beispielsweise Maschinen und deren Erhaltung). Demgemäß wird das Allgemeinvokabular für den MT relativ eingeschränkt sein.

[0078] Die technische Nomenklatur weist Wörter und Phrasen mit technischem Inhalt und Vokabular auf, das spezifisch für die Benutzeranwendung ist. Eintragungen mit technischem Inhalt sind Wörter und Phrasen, die spezifisch oder besonders für ein Betätigungsfeld oder eine Domäne sind. Die meisten technischen Wörter sind Hauptwörter, und zwar verwendet zur Benennung von Punkten, wie beispielsweise Teile, Komponenten, Maschinen oder Materialien. Sie können jedoch auch andere Klassen von Worten umfassen, wie beispielsweise Verbe, Adjektive und Adverbe. Offensichtlich, da diese Wörter nicht in der allgemeinen, täglichen Konversation verwendet werden, stehen sie im Kontrast zu Worten mit allgemeinem Inhalt.

[0079] Phrasen bzw. Ausdrücke mit technischem Inhalt sind Sequenzen mehrerer Worte, die aus all den vorhergehenden Klassen aufgebaut sind. Diese Phrasen sind die charakteristischste Form des Vokabulars technischer Dokumentation. Das Vokabular, das spezifisch für die Benutzeranwendung ist, ist der Teil der Terminologie, der abgegrenzte bzw. hervorgehobene Wörter und komplexe Begriffe enthält, die zur Benutzeranwendung erzeugt wurden. Diese umfassen die folgenden: Produktnamen, Dokumententitel, Akronyme, die von Benutzern verwendet werden, und Formennummern.

[0080] Die Entwicklung eines brauchbaren und vollständigen Vokabulars ist wichtig bei jeder Dokumentationsaufwendung. Wenn Dokumentation anschließend übersetzt wird, wird das Vokabular eine wichtige Quelle für den Übersetzungsaufwand. Der MT **120** ist so konstruiert, daß er die meisten funktionalen Eintragungen handhabt, die in Englisch verfügbar sind, mit der Ausnahme von jenen, die sich auf sehr persönliche (I, me, my, usw.), oder geschlechtsbasierende (hers, she, usw.), oder Pronomen (it, them, usw.) Verwendung beziehen. Dies wird eine Anzahl von technischen "Ausleihen" aus englischen Allgemeinworten umfassen (wie beispielsweise "truck" oder "length"). Die große Mehrheit des Vokabulars der eingeschränkten Sprache wird dann aus "Spezial-"Begriffen (d. h. technischen Begriffen) aus einem oder mehreren Worten bestehen, welche die Objekte bzw. Gegenstände und Prozesse der speziellen Domäne ausdrücken. In dem Ausmaß, in dem ein Vokabular in der Lage ist, den gesamten Notationsbereich über die spezielle Domäne auszudrücken, wird das Vokabular als vollständig bezeichnet.

[0081] Die Entwicklung eines auf eine Linie gebrachten, jedoch vollständigen Vokabulars trägt stark zum Erfolg des IATS-Systems **105** bei. Die eingeschränkte Sprache wird, durch Spezifizierung der richtigen und nicht richtigen Verwendung des Vokabulars sicherstellen, daß die Dokumente auf eine Weise erzeugt werden können, die zugänglich ist für eine schnelle, genaue und hochqualitative Maschinenübersetzung.

[0082] Vokabularteile bzw. -eintragungen sollten klare Ideen reflektieren und geeignet für die Zielleserschaft sein. Begriffe, die sexistisch, umgangssprachlich, idiomatisch, übermäßig kompliziert oder technisch, obskur bzw. unklar sind, oder welche auf einem anderen Weg eine Kommunikation verhindern, sollten vermieden werden. Diese und andere im allgemeinen akzeptierten stilistischen Betrachtungen sind nichts desto weniger wichtige Richtlinien für eine Dokumentenerzeugung im allgemeinen, auch wenn sie nicht notwendigerweise

verpflichtend für eine MT-orientierte Verarbeitung sind.

[0083] Es sei bemerkt, daß obwohl der Schwerpunkt bzw. das Volumen der Diskussion in diesem Dokument bezüglich der eingeschränkten Quellensprache und/oder Sprache im allgemeinen sich um Amerikanisches Englisch dreht, analoge Vergleiche in Verbindung mit allen anderen Sprachen angestellt werden können. Dem System **100** ist nichts inhärent, daß hier beschrieben wurde, daß es erforderlich macht, daß Amerikanisches Englisch die Quellensprache ist. Tatsächlich ist das System **100** nicht dazu aufgebaut, um nur mit Amerikanischem Englisch allein als Quellensprache zu arbeiten. Jedoch müssen die Datenbasen (d. h. das Domänenmodell), die mit dem LE **130** und dem MT **120** wechselwirken, ausgetauscht werden, um den Einschränkungen der besonderen Quellensprache zu entsprechen.

[0084] Den Regeln der Standardorthographie des Amerikanischen Englisch muß gefolgt werden. Nichtstandardbuchstabierungen, wie beispielsweise "thru" für "through", "moulding" für "molding" oder "hodometer" für "odometer" müssen vermieden werden. Großgeschriebene Worte (beispielsweise On-Off, Value Planned Repair) sollten nur zur Anzeige einer speziellen Bedeutung der Begriffe verwendet werden. Diese Begriffe müssen im Benutzeranwendungsvokabular aufgeführt werden. Das Gleiche ist der Fall für nichtstandardisierte Verwendung von Großschreibung (BrakeSaver). Gleichfalls müssen Abkürzungen, sofern sie verwendet werden (ROPS, API, PIN) im spezifischen Benutzeranwendungsvokabular aufgeführt sein. Das Format von Zahlen, Meßeinheiten und Daten muß konsistent sein.

[0085] Wiederverwendete Eintragungen der eingeschränkten Sprache sollten ebenso gemäß ihrer Bedeutung in der eingeschränkten Sprache verwendet werden. Wenn dies gemacht wird, stellt der Schreiber sicher, daß der MT immer ein Wort unter Verwendung des richtigen Wortsinns in der eingeschränkten Sprache übersetzt. Einige englische Worte können auch zu mehreren als einer syntaktischen Kategorie gehören. In der eingeschränkten Sprache sollten syntaktisch zweideutige Worte in Konstruktionen verwendet werden, die sie von Zweideutigkeiten befreien.

[0086] Ein schwieriges Problem, das von der speziellen Natur der Domäne bzw. des Bereichs herrührt, ist in einigen Gebieten die häufige Verwendung von langen zusammengesetzten Hauptwörtern. Die Modifikationsbeziehungen, die bei solchen zusammengesetzten Hauptwörtern vorliegen, werden in unterschiedlichen Sprachen unterschiedlich ausgedrückt. Da es nicht immer durchführbar ist, diese Beziehungen aus dem Quellentext wieder herzustellen und sie in der Zielsprache auszudrücken, können komplexe zusammengesetzte Hauptwörter mit den folgenden Charakteristika im MT-Lexikon aufgeführt sein:

– Technische Begriffe aus dem spezifischen Benutzeranwendungsvokabular; und
– Zusammengesetzte Begriffe, die aus mehr als einem Wort bestehen.

[0087] Komplizierte Hauptwort-Hauptwort-Verbindungen sollten soweit möglich vermieden werden. Jedoch ist mit einigen Eintragungen, die im Lexikon aufgeführt sind, der MT in der Lage, diese wichtige Charakteristik von Dokumentation handzuhaben. Man bemerke, daß eine Hauptwort-Hauptwortverbindung, die ein weit verbreitetes Merkmal der englischen Sprache ist, nicht notwendigerweise ein verbreitetes Merkmal anderer Sprachen sein muß, und demgemäß unterscheiden sich die Einschränkungen, unter welchen die eingeschränkte Sprache erzeugt wird, bezüglich der besonderen Quellensprache, die verwendet wird.

[0088] Englisch ist sehr reich an Verb-Partikel-Kombinationen, wobei ein Verb mit einer Präposition, einem Adverb oder anderen Sprachteilen kombiniert wird. Da das Partikel oft vom Verb durch Objekte oder andere Ausdrücke bzw. Phrasen getrennt sein kann, ruft dies eine Komplexität und Zweideutigkeit bei der MT-Verarbeitung des Eingabetextes hervor. Demgemäß sollten Verb-Partikel-Kombinationen umgeschrieben werden, wo immer möglich. Dies kann für gewöhnlich durch die Verwendung eines Einzelwortverbs statt dessen erreicht werden. Beispielsweise verwende man:

– "must" oder "need" statt "have to";
– "consult" statt "refer to";
– "start the motor" anstatt von "turn the motor on";

[0089] Volle bzw. vollständige Begriffe und Ideen sollten wo immer möglich verwendet werden. Dies ist insbesondere wichtig, wenn Mißverständnisse auftreten können. Beispielsweise in der Phrase:
"Use a monkey wrench to loosen the bolt ..."
darf das Wort "wrench" nicht ausgelassen werden. Während die meisten technisch bewanderten Menschen die Bedeutung bzw. Implizierung ohne dieses Wort verstehen würden, muß es während des Übersetzungsprozesses explizit gemacht werden. Ein CTE-Text (constrained text editor text = eingeschränkter Texteditortext) muß Vokabular haben, das explizit ausgedrückt ist, wo immer möglich; Abkürzungen oder verkürzte Begriffe

sollten in lexikalisch vollständige Ausdrücke umgeschrieben werden.

[0090] Man betrachte ein weiteres Beispiel:
"If the electrolyte density indicates that ..."

[0091] Hier ist die Bedeutung expliziter und vollständiger, wenn die Idee vollständig ausgedrückt wird:
"If measurement of the electrolyte density indicates that ..."

[0092] Schließlich werden bei den folgenden Sätzen, bei denen Wörter oder Phrasen fehlen, die unterstrichenen Wörter zur Verfügung gestellt, um die Bedeutung expliziter zu machen:
Turn the start switch key to OFF and remove the key.
Pull the backrest (1) up, and move the backrest to the desired position.
Jump starting; make sure the machines do not touch each other.

[0093] Wenn solche "Lücken" gefüllt sind, ist die Idee vollständiger und eine Übersetzung mit Bedeutung durch das IATS **105** wird sicherer. Übersetzungsfehler aufgrund von Lücken sind ein verbreiteter Grund für eine Nacheditierung. Deshalb sind Lücken nicht erlaubt.

[0094] Umgangssprachliches oder gesprochenes Englisch zieht oft die Verwendung von sehr allgemeinen Worten vor. Dies kann manchmal in einen Grad der Unbestimmtheit resultieren, der während des Übersetzungsprozesses aufgelöst werden muß. Beispielsweise sind Worte wie "conditions, remove, facilities, procedure, go, do, is for, make, get", usw. korrekt aber unpräzise.

[0095] In einem Satz wie:
When the temperature reaches 32°F, you must take special precautions.
vermittelt das Wort "reaches" nicht, ob die Temperatur fällt ("dropping") oder steigt ("rising"); einer dieser zwei Begriffe wäre hier exakter und der Text wäre genauso lesbar.

[0096] Einige Sprachen machen Unterschiede, wo Englisch sie nicht immer macht; beispielsweise sagen wir "oil" für entweder eine Schmiermittelflüssigkeit oder für die Verwendung bei der Verbrennung; wir sagen "fuel", ob es nun Diesel ist oder nicht. Ähnlich, wenn das Wort "door" isoliert verwendet wird, ist es nicht immer möglich, zu sagen, welche Art von "door" gemeint ist. Eine "car door"? Eine "building door"? Eine "compartement door"? Bei anderen Sprachen kann es notwendig sein, diese Unterscheidungen zu treffen. Wo immer möglich sollten vollständige Begriffe im Englischen verwendet werden.

C. Domänenmodell

[0097] Wissensbasierte Maschinenübersetzung (KBMT = Knowledge-based Machine Translation) muß unterstützt werden durch Weltwissen und durch linguistisches, semantisches Wissen über Bedeutungen von lexikalischen Einheiten und ihren Kombinationen. Die KBMT-Wissensbasis muß in der Lage sein, nicht nur eine allgemeine, taxonomische Domäne von Objekttypen zu repräsentieren, wie beispielsweise "car is a kind of a vehicle" (ein Auto ist eine Art eines Fahrzeugs), "a door handle is a part of a door" (ein Türgriff ist ein Teil einer Tür); wobei Artefakte durch die Eigenschaft "made-by" (hergestellt aus) charakterisiert sind (neben anderen Eigenschaften); sie muß auch Wissen über besondere Beispiele von Objekttypen repräsentieren (beispielsweise kann "IBM" in das Domänenmodell als ein markiertes Beispiel des Objekttyps "Unternehmen" aufgenommen sein) sowie Beispiele von (potentiell komplexen) Ereignistypen (beispielsweise ist die Wahl von George Bush als Präsident der Vereinigten Staaten ein markiertes Beispiel der komplexen Aktion "wählen"). Der ontologische Teil der Wissensbasis nimmt die Form einer Multihierarchie von Konzepten an, die über Taxonomie aufbauende Verbindungen verbunden sind, wie beispielsweise "is-a" (ist ein), "part-off" (Teil von) und einigen anderen. Wir nennen die resultierende Struktur eine Multihierarchie, weil den Konzepten erlaubt wird, verschiedene Ursprünge bzw. Eltern auf jedem Verbindungstyp zu besitzen.

[0098] Das Domänenmodell oder Konzeptlexikon enthält ein ontologisches Modell, das gleichförmige Definitionen von Basiskategorien vorsieht (wie beispielsweise Objekte, Ereignistypen, Beziehungen, Eigenschaften, Episoden, usw.), die als Aufbaublöcke für Beschreibungen von besonderen Domänen verwendet werden. Dieses "Welt"-Modell ist relativ statisch und ist in einem mehrfach zwischenverbundenen Netzwerk von ontologischen Konzepten organisiert. Die allgemeine Entwicklung einer Ontologie einer Anwendungs-(Teil-)-Welt ist im Stand der Technik gut bekannt. Siehe beispielsweise Brachmann und Schmolze, An Overview of the KL-ONE Knowledge Representation System, Cognitive Science, vol. 9, 1985; Lenat, u. a., Cyc: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks, Al Magazine, VI: 65–85, 1985;

Hobbs, Overview of the Tacitus Project, Computational Linguistics, 12: 3, 1986; und Nirenburg u. a., Acquisition of Very Large Knowledge Bases: Methodology, Tools and Applications, Center for Machine Translation, Carnegie Mellon University (1988).

[0099] Die Ontologie ist eine sprachunabhängige konzeptmäßige Repräsentation einer spezifischen Teilwelt, wie beispielsweise Fehlerbeseitigung bei Schwergerät und Reparatur desselben oder die Wechselwirkung zwischen Personalcomputern und ihren Benutzern. Sie sieht eine semantische Information vor, die notwendig in der Teilsprachendomäne für die Satzteilbestimmung (englisch: parsing) des Quellentextes in einen Zwischensprachentext und Erzeugung von Zieltexten aus den Zwischensprachentexten ist. Das Domänenmodell muß ausreichend detailliert sein, um ausreichend semantische Einschränkungen vorzusehen, die Zweideutigkeiten bei der Satzteilbestimmung eliminieren, und das ontologische Modell muß gleichförmige Definitionen von grundlegenden ontologischen Kategorien vorsehen, die die Aufbaublöcke für Beschreibungen der besonderen Domänen sind. In einem Weltmodell können die ontologischen Konzepte zuerst unterteilt werden in Objekte, Ereignisse, Kräfte (eingeführt, um Urhebern ohne Absicht Rechnung zu tragen) und Eigenschaften. Eigenschaften können weiter unterteilt werden in Beziehungen und Attribute. Beziehungen werden als Abbildungen von Konzepten definiert (beispielsweise ist "belongs-to" (gehört zu) eine Beziehung, da es ein Objekt in den Satz abbildet {*Mensch *Organisation}), während Attribute als Abbildungen von Konzepten in speziell definierte Wertsätze definiert werden (beispielsweise ist "Temperatur" ein Attribut, das physikalische Objekte in Werte auf einer halboffenen Skala [0,*] abbildet, und zwar mit der Gradeinteilung auf der Kelvinskala). Konzepte werden typischerweise als Rahmen repräsentiert, deren Schlitze Eigenschaften sind, die im System vollständig definiert sind.

[0100] Domänenmodelle sind ein notwendiger Teil eines jeglichen wissensbasierenden Systems, nicht nur von einem wie eine wissensbasierende Maschinenübersetzung. Das Domänenmodell ist eine semantische Hierarchie von Konzepten, die in der Übersetzungsdomäne auftreten. Beispielsweise können wir das Objekt *O-VEHICLE so definieren, daß es *O-WHEELED-VEHICLE und *O-TRACKED-VEHICLE umfaßt, und wobei das Erstere *O-TRUCK, *O-WHEELED-TRACTOR, usw. umfaßt. Ganz unten von dieser Hierarchie sind die spezifischen Konzepte, die der Terminologie in CSL entsprechen. Wir nennen diesen unteren Teil bzw. Bodenteil die geteilte K/DM. Um genau zu übersetzen, müssen wir semantische Einschränkungen den Rollen auferlegen, die die verschiedenen Konzepte spielen. Beispielsweise die Tatsache, daß die Urheberrolle einer *E-DRIVE-Tätigkeit durch einen Menschen ausgefüllt werden muß, ist eine semantische Einschränkung, die dem *O-VEHICLE auferlegt wird, und automatisch von allen Arten des Fahrzeugs (Vehikels) geerbt wird (um somit Wiederholungsarbeit bei der Handkodierung eines jeden Beispiels einzusparen). Der Verfassungsteil des Domänenmodells erweitert das K/DM mit Synonymen, die nicht in CSL sind, und andere Information, um eine nützliche Rückkopplung für den Verfasser vorzusehen, wenn er oder sie jedes Informationselement zusammensetzen bzw. komponieren.

[0101] **Fig.** 5 stellt konzeptmäßig das Domänenmodell (DM) dar, das von der vorliegenden Erfindung verwendet wird. Das DM **500** ist eine Repräsentation eines erklärenden bzw. festlegenden Wissens über das CSL-Vokabular, das vom MT **120** und dem LE **130** verwendet wird. Das DM **500** ist aus drei unterschiedlichen Teilen gemacht:

1. Ein Kerndomänenmodell (K/DM) **510** enthält alle lexikalische Information, die sowohl vom MT-Analysierer **127** als auch dem LE **130** angefordert wird; insbesondere umfaßt der Kern bzw. Kernel alle CSL-lexikalischen Eintragungen (Worte und Phrasen) mit assoziierten semantischen Konzepten, Wortklassen, morphologischer Information, usw.
2. Ein MT-Domänenmodell (MT/DM) **520**, das Information enthält, die nur vom MT-Analysierer **127** angefordert wird. Das MT-Domänenmodell ist die Hierarchie der Konzepte, die für eine unzweideutige Abbildung und eine semantische Verifizierung bei der Übersetzung verwendet werden. Es umfaßt Auswahlseinschränkungen bezüglich der Konzepte und eine hierarchische Klassifizierung der Konzepte.
3. Ein LE-Domänenmodell (LE/DM) **530** enthält Information, die nur vom LE **130** angefordert wird; dies umfaßt nicht CSL-Synonyme für CSL-lexikalische Eintragungen, Wortbuchdefinitionen von CSL-lexikalischen Eintragungen und Beispiele für die Verwendung von CSL-lexikalischen Eintragungen.

[0102] Das Kern/DM **510** wird eine lexikalischen Eintragung für jede CSL-lexikalische Eintragung bzw. Punkt (Wort oder Phrase) enthalten. (Eine "lexikalische Eintragung" besteht aus einem lexikalischen Punkt – ein Wort oder eine Phrase – und minimaler Weise seinem assoziierten semantischen Konzept und seiner Wortklasse, beispielsweise wenn das Wort "leak" in CSL sowohl ein Hauptwort als auch ein Verb ist, würde es zwei lexikalische Eintragungen haben.) Jeder lexikalische Punkt wird mit zusätzlicher Information aktualisiert, die vom LE **130** und/oder dem MT **120** angefordert wird, wie beispielsweise eine Definition und unregelmäßige morphologische Varianten.

[0103] Das geteilte K/DM **510** beschleunigt die Verfeinerung und Erweiterung der CSL, erspart Verdoppelungen vom Aufwand bei Verfassungs- und Übersetzungskomponenten und sieht durch den Menschen lesbare Strukturen vor, um eine Wartung und Erweiterungen zu ermöglichen.

[0104] Das K/DM **510** ist ein Lexikon, das sowohl die syntaktische als auch die semantische Information über Begriffe (Wörter und Phrasen) im eingeschränkten Sprachtext enthält. Es ist die zentrale lexikalische Wissensquelle für die Analyseseite des automatisierten Maschinenübersetzungs-(MT)-Prozesses. Das K/DM **510** wird ebenso als Basis für die LE/DM verwendet.

[0105] Das K/DM **510** umfaßt eine getrennte Eintragung für jeden Begriff in jeder syntaktischen Kategorie. (Somit gibt es für ein Wort wie "truck", welches sowohl ein Hauptwort bzw. Substantiv und ein Verb ist, zwei Eintragungen.) K/DM-Eintragungen enthalten die folgende Information:

– Wurzel (beispielsweise "truck");
– Redeteil bzw. Wortklasse (beispielsweise N);
– für Inhaltsworte das Konzept oder die Bedeutung (beispielsweise O-TRUCK);
– morphologische Information (beispielsweise unregelmäßige Flexion bzw. Beugung);
– syntaktische Information (beispielsweise ob ein Hauptwort zählbar oder Masse ist);
– Definitionsinformation: Kurze Definitionen und Textbeispiele, die verschiedene Sinngebungen und Verwendungen der Worte dokumentieren, und eine Spezifikation des Sinns, in welchem das Wort in der eingeschränkten Sprache verwendet werden soll.

[0106] Das DM **500** wird in drei Sätzen von externen, durch den Menschen lesbaren Dateien definiert, die durch den Prozeß bzw. die Prozesse gelesen werden können, die ihre Verwendung anfordern. Da der MT **120** und der LE **130** in getrennten Prozessen laufen werden, ist die Information im Modell intern in zwei Formen repräsentiert: Eine für die Teile des DM, die durch den MT **120** angefordert wird, und eine weitere für die Teile, die durch den LE **130** angefordert werden. So wird das K/DM **510** in einen Satz von Dateien definiert, die in beiden Formen repräsentiert sein können; das LE/DM **530** ist nur in der Form repräsentiert, wie sie durch den LE **130** verwendet wird; und das MT/DM **520** ist nur in der Form repräsentiert, wie sie vom MT **120** verwendet wird. Weiter unten werden die externen Dateiformate beschrieben, ferner der Inhalt der verschiedenen Teile des DM und die interne Repräsentation der Information, wie sie vom LE **130** verwendet wird.

[0107] Zur Wiederholung, das K/DM enthält alle Information, die sowohl vom MT **120** als auch vom LE **130** angefordert wird. Dies umfaßt einen CSL-lexikalischen Punkt – das Basiswort, die Phrase oder einen zitierten Begriff und ein semantisches Konzept – wobei das semantische Konzept, daß mit dem lexikalischen Punkt bzw. der Eintragung assoziiert ist, in einer lexikalischen Eintragung durch einen "Konzeptnamen" repräsentiert ist. Ferner umfaßt es einen Redeteil bzw. eine Wortklasse – und zwar einen aus einem festen Satz von Redeteilen (d. h. Verb, Adjektiv, usw.) eine Definition – eine grobe Definition für allgemeine Vokabelbegriffe, um klarzustellen, welche der vielen Sinngebungen ein CSL-lexikalischer Punkt haben kann, und unregelmäßige morphologische Varianten – eine Liste von unregelmäßigen morphologischen Formen und für jede den Namen der morphologischen Transformation. Beispiele für Namen für morphologische Transformationen von Verben sind "Vergangenheit", "dritte Person Einzahl Gegenwart", "Partizip Perfekt", "Partizip Präsens". Der Wert dieses Feldes für das Wort "drive" würde beispielsweise sein ((Vergangenheit drove) (Partizip Perfekt driven)), was anzeigt, daß diese beiden Formen des Verbs unregelmäßig sind und alle anderen Formen regelmäßig. Schließlich umfaßt das K/DM typographische Einschränkungen – beispielsweise wenn der lexikalische Punkt bzw. die Eintragung in Großbuchstaben sein muß, der erste Buchstabe groß geschrieben werden muß, usw.

[0108] Das MT/DM **520** enthält Information, die nur durch den MT **120** angefordert wird. Dies umfaßt: Auswahlseinschränkungen bezüglich der Konzepte und hierarchische Klassifikation der Konzepte für die Organisation und die Vererbung von Auswahlseinschränkungen.

[0109] Das LE/DM **530** wird Nicht-CSL-Synonyme enthalten, um den Verfasser dabei zu unterstützen, gültige CSL-lexikalische Eintragungen auszuwählen. Zusammen werden der Kern und das LE/DM alle Information und alle Einschränkungen enthalten, die zur Charakterisierung des CSL-Lexikons erforderlich sind, und zwar mit Unterstützung der LE-Vokabularüberprüfung (weiter unten beschrieben). Das LE/DM enthält zusätzliche Information, die nur vom LE-Vokabelüberprüfer bzw. der LE-Vokabularüberprüfung angefordert wird.

[0110] Dies umfaßt: Eine Wörterbuchdefinition – die Definition des Wortes oder der Phrase, die dem Verfasser durch den LE präsentiert wird, Nicht-CSL-Synonyme – Synonyme für die CSL-lexikalischen Eintragungen, die der Verfasser beim Schreiben der Dokumente verwenden könnte, und ein Verwendungsbeispiel – ein Beispiel des Worts oder der Phrase in einem CSL-Satz, und zwar zur Präsentation für die Verfasser durch den LE.

[0111] Der Zweck des Einschließens dieser Information in das LE/DM ist es, die Verfasser dabei zu unterstützen, das sie sicherstellen, daß ihr Geschriebenes aus gültigen CSL-Worten und -Phrasen zusammengesetzt ist. Die Wörterbuchdefinitionen und Verwendungsbeispiele werden die Verfasser dabei unterstützen, sicherzustellen, daß sie ein Wort oder eine Phrase einer Wortklasse bzw. eines Redeteils verwenden und mit einer Bedeutung, die in CSL erlaubt ist; jedoch sind Wörterbuchdefinitionen oder Verwendungsbeispiele nicht für jede CSL-lexikalische Eintragung erforderlich. Statt dessen werden sie nur für den kleinen Prozentsatz von zweideutigen oder vagen Begriffen erforderlich sein, deren CSL-Bedeutung nicht sofort für den Verfasser klar ist. Dies beträgt wahrscheinlich weniger als die Hälfte der lexikalischen Eintragungen bzw. Punkte im DM. Beispielsweise werden Funktionsworte wie "for" und "the" keine Definitionen oder Beispiele erfordern; viele technischen Begriffe, insbesondere jene mit sehr spezifischen technischen Bedeutungen, brauchen keine Definitionen oder auch Beispiele erforderlich machen.

[0112] Die Nicht-CSL-Synonyme im LE/DM werden die Verfasser, die ein Nicht-CSL-Wort oder -Phrase schreiben, dabei unterstützen, ein synonymes oder verwandtes CSL-Wort oder eine CSL-Phrase auszuwählen, mit der es ersetzt werden soll. Es ist wünschenswert, daß der Vokabelüberprüfer Information nicht nur über Synonyme vorsieht, die in der selben Wortklasse wie das Nicht-CSL-Wort sind, mit dem sie synonym sind, sondern auch über verwandte Worte, die den Autor bei der Wortumgestaltung des Satzes helfen. Wenn die letzteren umfaßt sind, muß das LE/DM Information über diese verwandten Worte zusätzlich zum obligatorischen Inhalt enthalten.

D. Spracheditor

[0113] Unter Bezugnahme auf **Fig.** 1(b) ist der eingeschränkte Spracheditor (LE = language editor) **130** ein Satz von Werkzeugen zur Unterstützung der Verfasser und Redakteure bzw. Editoren beim Erzeugen von Dokumenten innerhalb der Grenzen der CSL. Diese Werkzeuge werden einen Verfasser bei der Verwendung des geeigneten bzw. richtigen CSL-Vokabulars und seiner Grammatik zum Schreiben von Bedienungsdokumentation bzw. Dienstdokumentation unterstützen. Der LE **130** ist als eine "Erweiterung" des SGML-Texteditors **140** aufgebaut. Obwohl der LE **130** die selben Kommunikationskanäle wie der SGML-Texteditor **140** verwendet, sind die Funktionen der beiden gegenseitig ausgeschlossen. Jedoch ist die Benutzerschnittstelle, die für die Wechselwirkung mit dem LE **130** verwendet wird, eine "nahtlose Erweiterung" der SGML-Texteditorschnittstelle.

[0114] Der Verfasser **160** erzeugt Dokumente im SGML-Texteditor **140** und ruft den LE **130** auf. Der LE **130** informiert den Verfasser, ob individuelle Wörter in einem Dokument nicht-CSL sind, und er wird in der Lage sein, Synonyme in der CSL für Wörter vorzuschlagen, die relevant für die Informationsdomäne der Benutzeranwendung sind, jedoch nicht in der CSL sind. Zusätzlich teilt der LE **130** dem Verfasser mit, ob oder ob nicht der Text in einer Datei CSL-syntaktische Einschränkungen erfüllt.

[0115] Die LE **130**-Software umfaßt folgendes: Einen Vokabularüberprüfer bzw. eine Vokabularüberprüfung, einen Grammatiküberprüfer bzw. -überprüfung, einschließlich einer Schnittstelle durch den MT-syntaktischen Analysierer, der die Kerngrammatiküberprüfungsfunktionalität vorsieht, und eine Benutzerschnittstelle bzw. ein Benutzerinterface (UI = User Interface). Zusätzlich wird die CSL-Vokabularinformation, die vom CSL-LE verwendet wird, im K/DM und im LE/DM repräsentiert sein.

[0116] Der LE **130** wird bestätigen, daß alles Vokabular und alle Satzstrukturen in einem Dokument konform mit der CSL-Spezifikation sind. Der LE **130** markiert das Dokument mit einer SGML-Kennzeichnung, die diese CSL-Bestätigung repräsentiert. Eine Überprüfung muß auf den ganzen Text in einem Dokument angewandt bzw. durchgeführt werden, was folgendes umfaßt: Sätze, Überschriften, Listeneintragungen, Untertitel, Beschriftungen in Graphiken und Information in Tabellen.

[0117] Da die vorliegende Erfindung auf der Prämisse basiert, daß Verfasser so produktiv wie möglich während einer CSL-Überprüfungssitzung sein sollten, und daß die Verfasser nicht an mehreren Verfassungsdokumenten auf einmal arbeiten sollten, ist ein Stapelverarbeitungsbetriebsmodus, der es erforderlich macht, daß ein Benutzer ein Dokument zur Bearbeitung eingibt und wartet bis das gesamte Dokumente fertiggestellt ist, bevor er oder sie irgendeine Rückkopplung erhält, nicht geeignet. Der LE **130** sieht einen interaktiven Betriebsmodus für die Vokabularüberprüfung, Grammatiküberprüfung und die interaktive Zweideutigkeitsentfernung vor.

[0118] **Fig.** 6 zeigt ein Hochebenenfließdiagramm des Betriebs des LE **130**. Der LE **130** nimmt als Eingabe einen Text **605**, der zweideutig und uneingeschränkt sein kann. Der potentiell zweideutige, uneingeschränkte

Eingabetext **605** wird zuerst mit einer Vokabelüberprüfung **610** überprüft, die ihre Funktionen (wie weiter unten beschrieben) mit der Hilfe einer Buchstabierüberprüfung **615** durchführt. (Die Dienste der Buchstabierüberprüfung werden in diesem Ausführungsbeispiel durch die Buchstabierüberprüfung bereitgestellt, die regelmäßig durch den Host-TE **140** dargestellt wird.) Sobald die Vokabelüberprüfung **610** abgeschlossen ist, überprüft sie alle notwendigen Korrekturen und bringt diese an (mit der Hilfe des Verfassers), dann wird der lexikalisch eingeschränkte Text **617** an eine Grammatiküberprüfung **620** geliefert. Die Grammatiküberprüfung **620** erzeugt einen syntaktisch korrekten CSL-Text **625**. Der eingeschränkte syntaktisch korrekte Text **625** wird von Zweideutigkeiten befreit, wie im Block **630** gezeigt. Das Ergebnis der Zweideutigkeitsentfernung ist ein übersetzbarer, nicht zweideutiger eingeschränkter Text **635**. Der übersetzbare Text **635** kann in eine Fremdsprache übersetzt werden, ohne eine Nacheditierung zu erfordern. Die Genauigkeit der resultierenden Übersetzung macht ebenso eine Nacheditierung nicht notwendig.

1. Vokabularüberprüfung

[0119] **Fig.** 7 zeigt ein Fließdiagramm des Betriebs der Vokabelüberprüfung **610**. Die Vokabelüberprüfung **610** identifiziert Worte, die nicht in der CSL bekannt sind. Die Vokabularüberprüfung **610** identifiziert das Auftreten von Nicht-CSL-Worten, und zwar im Text des Verfassers, und unterstützt einen Verfasser beim Auffinden von gültigen CSL-Ersetzungen für Nicht-CSL-Worte. Sie erkennt Wortgrenzen in einem Dokument und identifiziert jedes Auftreten bzw. Beispiel einer lexikalischen Einheit bzw. eines lexikalischen Punktes, der in CSL nicht bekannt ist.

[0120] Wie in Block **706** gezeigt, wird der erste Begriff einer Einheit zur Überprüfung ausgewählt. Der Begriff wird dann überprüft, wie in Block **710** gezeigt, und zwar gegen eine CSL-lexikalische Datenbasis (beispielsweise ein Wörterbuch), das alle CSL-Worte enthält. Wenn der Begriff im CSL-Wörterbuch nicht gefunden wird, wird der Begriff bezüglich seiner Buchstabierung gegen ein Standardwörterbuch überprüft, wie in Block **722** gezeigt. Wenn das Wort falsch buchstabiert wurde, wird der Verfasser mit Mitteln zur Korrektur des Buchstabierfehlers versehen (beispielsweise stellt die Vokabularüberprüfung **610** Buchstabieralternativen dar), wie in Block **726** gezeigt.

[0121] Die Eintragung bzw. der Punkt wird dann überprüft, um zu bestimmen, ob er sich im CSL-Vokabular befindet, wie im Block **734** gezeigt. Wenn der Punkt im CSL-Vokabular ist, dann schreitet die Prozedur fort zu Block **718**. Wenn jedoch der Punkt nicht im CSL-Vokabular ist, überprüft das System, um herauszufinden, ob das LE/DM ein Synonym für den zu überprüfenden Punkt enthält, wie in Block **736** gezeigt. Wenn zumindest ein Synonym im LE/DM existiert, stellt das System das Synonym bzw. die Synonyme dar, die Teil des CSL-Vokabulars sind, und erlaubt dem Verfasser, eine Auswahl zu treffen, wie im Block **738** gezeigt. Sollte das LE/DM jedoch nicht über ein Synonym für den Punkt verfügen, der überprüft wird, dann hat der Verfasser die Gelegenheit, seine bzw. ihre Eingabe zu überarbeiten, wie im Block **740** gezeigt. Das Ergebnis dieser Überarbeitung geht zurück zu Block **710**. Sobald eine legale Auswahl durch den Verfasser getroffen wurde, schreitet die Prozedur **700** dann zu Block **718** fort.

[0122] Wenn ein Nicht-CSL-Wort identifiziert wird, hat der Verfasser die folgenden Optionen: Er oder sie kann eine Alternative von einem Ersatz für das Wort im Dokument auswählen, oder er bzw. sie kann einen neuen Punkt bzw. eine Eintragung eingeben und diesen für das Wort im Dokument einsetzen bzw. ersetzen, typischerweise wählt der Verfasser eines der Synonyme für die Ersetzung des Nicht-CSL-Punktes aus: Wenn sich der Verfasser entschließen sollte, dieses Problem zu überspringen, würde das Fehlen der Auflösung in einem Versagen beim Genehmigen des Textes als CSL resultieren.

[0123] Block **718** überprüft, um festzustellen, ob es noch weitere Begriffe in der Einheit gibt. Wenn es keine weiteren Begriffe gibt, hält die Prozedur **700** an. Ansonsten wird der nächste Begriff ausgewählt, wie im Block **714** gezeigt, und die Prozedur **700** beginnt wieder vom Block **710** aus.

[0124] Insbesondere identifiziert die Vokabularüberprüfung **610** jedes Auftreten bzw. jedes Beispiel eines lexikalischen Punktes, der nicht in der CSL bekannt ist. Für jedes solche Wort wird die Vokabelüberprüfung **610** bestimmen, welche der folgenden Beschreibungen anwendbar ist, und sie wird unterstützende Information an die Benutzerschnittstelle berichten, wie weiter unten aufgeführt:

– ein Nicht-CSL-Wort mit bekannten CSL-Synonymen; in diesem Fall wird die Vokabularüberprüfung **610** die Synonyme identifizieren. Beispielsweise sei angenommen, daß das Wort "let" Nicht-CSL ist –

[0125] Die Eingabe des Verfassers, sobald überprüft: Open the valve and let more nitrogen go to the accumulator.

[0126] VC-Nachricht: Der Begriff ist Nicht-CSL, aber es gibt verwandte CSL-Alternativen.

[0127] CSL-Alternativen: allow, allowed, enable, enabled, permit, permitted, leave, left

[0128] CSL-Satz wie editiert: Open the valve and allow more nitrogen to go to the accumulator.
– ein Wort, das in CSL nur als Teil einer Phrase erscheinen kann, jedoch nicht in einer CSL-Phrase im aktuellen Kontext verwendet wird; in diesem Fall wird die Vokabularüberprüfung (Vocabulary Checker = VC) **610** akzeptierbare CSL-Phrasen berichten, die das Wort enthalten –

[0129] Eingabe des Verfassers, wenn überprüft: The first time the valve lash is checked, the injector timin should be checked.

[0130] VC-Nachricht: Der Begriff wird in einem Nicht-CSL-Kontext verwendet.

[0131] CSL-Alternativen: advance signal timing, advance timing groove, timing gear, timing mechanism

[0132] CSL-Satz, wie editiert: The first time the valve lash is checked, the injector timing mechanism should be checked.
– ein Wort oder eine Phrase, die in CSL innerhalb von doppelten Anführungszeichen erscheinen müssen, jedoch im aktuellen Text nicht von Anführungszeichen eingeschlossen sind; in diesem Fall wird die Vokabularüberprüfung **610** berichten, daß der Begriff in Anführungszeichen gesetzt werden sollte –

[0133] Eingabe des Verfassers, wenn überprüft: For more details, read the Testing and Adjusting article in the next section.

[0134] VC-Nachricht: Dieser Begriff ist im allgemeinen in Anführungszeichen eingeschlossen.

[0135] CSL-Alternative: Keine CSL-Satz, wenn editiert bzw. redigiert: For more details, read the "Testing and Adjusting" article in the next section.
– ein Wort oder eine Phrase, die in CSL mit spezifischer, obligatorischer Großschreibung erscheinen muß, denen diese Großschreibung im aktuellen Kontext jedoch fehlt (beispielsweise ein Akronym, das in Kleinschreibung dargestellt ist); in diesem Fall wird die Vokabularüberprüfung **610** die korrekte CSL-Form bzw. – Formen berichten –

[0136] Eingabe des Autors, wenn überprüft: Turn the screw until the pressure gauge reads 0 kpa (0 psi).

[0137] VC-Nachricht: Der Begriff ist nicht richtig großgeschrieben.

[0138] CSL-Alternative: kPa

[0139] CSL-Satz wie editiert: Turn the screw until the pressure gauge reads 0 kPa (psi).
– ein Nicht-Wort (das ist eine Gruppe von Buchstaben, die ein falsch buchstabiertes Wort repräsentieren), das bekannte Buchstabieralternativen besitzt; in diesem Fall wird die Vokabularüberprüfung **610** die Buchstabieralternativen identifizieren, und zwar unabhängig, ob das Resultat in CSL ist (der Benutzer wird die gewählte Alternative für eine weitere Überprüfung wiedereingeben) –

[0140] Eingabe des Verfassers, wenn überprüft: When it is necesary to raise the boom, the boom must have correct support.

[0141] VC-Nachricht: Der Begriff ist Nicht-CSL.

[0142] CSL-Alternative: necessary

[0143] CSL-Satz wie editiert: When it is necessary to raise the boom, the boom must have correct support.
– ein Wort, das nicht in CSL ist und über das das System nichts weiß. Die Nachricht für ein unbekanntes Wort oder eine Phrase gibt dem Verfasser die Gelegenheit, die Wortgebung gesamt zu verändern oder den nichtzugelassenen Ausdruck von der Überprüfung abzuschirmen, je nachdem, wie es der Fall erfordert. Beim folgenden Beispiel verwendet der Verfasser eine SGML-Kennzeichnung, um dem System mitzuteilen, die verletzende Sprache zu übersehen und sie intakt zu belassen –

[0144] Eingabe des Verfassers, wenn überprüft: Put approximately 0.9 L (1 quart) of SAE10W hydraulic oil in the nitrogen end of the accumulator.

[0145] VC-Nachricht: Der Begriff ist unbekannt.

[0146] CSL-Alternative: keine

[0147] CSL-Satz, wie editiert: Put approximately 0.9 L (1 quart) of <sic>SAE10W</sic> hydraulic oil in the nitrogen end of the accumulator.

– eine Satzzeichenmarkierung oder ein spezielles Symbol, das in CSL oder irgendeinem Kontext nicht erlaubt ist

[0148] In Fällen, in denen ein Nicht-CSL-Wort keine direkten CSL-Synonyme besitzt (das sind Worte, die es direkt in einem Dokument ersetzen könnten), kann das System verwandte CSL-Worte oder Phrasen identifizieren, die ein Verfasser verwenden könnte, um die beabsichtigte Idee auszudrücken. Diese Funktionalität versieht den Verfasser mit einer zusätzlichen Unterstützung bei der Wortänderung eines Satzes, um nur CSL-Vokabular einzuschließen. Jedoch könnten Veränderungen zur Verwendung dieser verwandten Worte nicht mit der automatischen Ersetzungsmöglichkeit vervollständigt werden, die für Synonyme vorgesehen ist, da die Veränderungen einige Modifikationen bezüglich der Satzstruktur erforderlich machen würden. Beispielsweise, wenn "can" in CSL ist und "capable" nicht, würde einem Verfasser, der den folgenden Satz schrieb
The system is capable of being programmed for several customerspecified parameters.
mitgeteilt werden, daß "capable" [[capable]] nicht ein CSL-Wort war. Obwohl das Wort "can" [[can]] in CSL ist, kann weder das Wort "capable" noch die Phrase "is capable of" [["is capable of"]] direkt durch "can" ersetzt werden, ohne weitere Veränderungen am Satz zu benötigen.

2. Grammatiküberprüfung

[0149] Der Zweck der Grammatiküberprüfung bzw. des Grammatiküberprüfers ist es, Stellen zu identifizieren, wo ein Text des Verfassers nicht konform mit den CSL-grammatikalischen Einschränkungen ist, und die Aufmerksamkeit des Verfassers auf diese Stellen zu fokussieren bzw. zu lenken. Die Funktionalität der Grammatiküberprüfung **620** wird durch ein Analysemodul **127** des MT-Systems **120** vorgesehen, und zwar erweitert, um dem System zu erlauben, das Auftreten bzw. Beispiele von syntaktischen und semantischen Zweideutigkeiten zu berichten. Die Grammatiküberprüfungsschnittstelle erlaubt dem Verfasser, interaktiv auf Anforderungen zur Klärung von Zweideutigkeiten zu antworten. Es ist möglich, daß ein Satz in einer eingeschränkten Sprache sein kann, das er jedoch mehr als eine Interpretation besitzen kann. Die Grammatiküberprüfungsschnittstelle wird eine Anzeige von den zwei oder mehr möglichen Bedeutungen des Satzes für den Verfasser vorlegen und eine Klärung anfordern. Ein Beispiel eines zweideutigen Satzes wäre: "Check the cylinders on the inside." Sind die "cylinders" an der "inside" (Innenseite) angeordnet oder soll man "the inside" (die Innenseite) der "cylinders" (Zylinder) überprüfen? Es gibt zwei Arten von möglichen Zweideutigkeiten:
Lexikalische Zweideutigkeiten. Lexikalische Zweideutigkeiten treten auf, wo ein Wort mehr als zwei Bedeutungen in der eingeschränkten Sprache haben kann. Während es wünschenswert ist, das in der eingeschränkten Sprache jedes Wort nur eine Bedeutung pro Wortklasse bzw. Redeteil hat, gibt es einige Worte, die mehr als eine Bedeutung haben. Beispielsweise kann das Wort "gas" die Bedeutung "natural gas" oder "gasoline" haben.

[0150] Auf der lexikalischen Ebene können ebenso Probleme durch ein Wort hervorgerufen werden, welches in zwei verschiedenen syntaktischen Rollen in CSL verwendet werden kann. So ist es der Fall mit "fuel", das entweder ein Hauptwort oder ein Verb in CSL sein kann. Wenn der Autor bzw. Verfasser einen Satz eingibt, bei dem die syntaktische Rolle nicht klar ist, kann die Grammatiküberprüfung (GC = Grammar Checker) **620** dem Verfasser wie folgt antworten bzw. entgegnen:
Eingabe des Verfassers, wenn überprüft: The sensor is attached to fuel rack.
GC-Nachricht: Der Begriff kann als ein Hauptwort oder als ein Verb verwendet werden.

[0151] An diesem Punkt hat der Verfasser die Option, den Satz ohne die Hilfe des Systems zu editieren bzw. redigieren (was einfach das Umschreiben und Neueingeben in die Überprüfung erfordert). Wenn der Verfasser eine Hilfeaufforderung auswählt, kann das System spezifische Instruktionen zur Behandlung von Problemen des selben Typs anbieten. In diesem Fall ist die Hilfe spezifisch:
Hilfe!
GC-Nachricht: Wenn das Wort ein Hauptwort ist, kann man ein Bestimmungswort davor setzen, wenn es ein Verb ist, kann ein Bestimmungswort danach hilfreich sein?

Beispiel: The ship sinks vs. Ship the sinks.

[0152] Der Verfasser fährt dann damit fort, den Satz zu editieren und gibt ihn wieder in die Grammatiküberprüfung **620** ein.

[0153] Strukturelle Zweideutigkeit. Strukturelle Zweideutigkeit tritt auf, wo Worte in einem Satz gruppiert werden können und zwar auf mehr als nur einem Weg. Beispielsweise: "Remove the valve with the lever." Bildet die Phrase "with the lever" eine Einheit mit der Phrase "the valve", oder bildet sie statt dessen eine Einheit mit dem Verb "remove"? In anderen Worten ist dies ein Satz über ein "valve" (Ventil), das einen "lever" (Hebel) daran angebracht hat, oder ist es über die Verwendung eines "lever" (Hebels) "to remove a valve" (zum Entfernen eines Ventils)?

[0154] Im IATS **105** ist die Komponente, die zur Beantwortung dieser Frage konstruiert ist, das Domänenmodell **137**, das in einer solchen Weise aufgebaut ist, das es das Auftreten von solchen Zweideutigkeiten minimiert.

[0155] Wie in **Fig.** 5 gezeigt, enthält das DM/MT **520**, das ausschließlich den Maschinenübersetzungsprozeß unterstützt, zwei Typen von Information. Einerseits unterstützt die semantische Information (A) die Identifizierung von Beziehungen zwischen Konzepten. Andererseits spezifiziert die Kontextinformation (B) für ein besonderes Verb die sogenannten Tiefenfälle oder Argumente, die ein solches Verb annehmen kann. Im betrachteten Beispiel sei zuerst betrachtet, wie die semantische Information (A) und die Kontextinformation (B) den Analysierer **127** dabei unterstützen, die grammatikalische Struktur von "Remove the valve with the lever" zu bestimmen.

[0156] Unter vielen semantischen Beziehungen gibt es eine Beziehung "is a part of", die beispielsweise zwischen dem Konzept "hat" und dem Konzept "costume" in Kraft tritt, wobei der "hat" (Hut) "is a part of" (ein Teil ist von) des "costume" (Kostüms). Die selbe Beziehung tritt in Kraft zwischen dem Konzept "sole" (Sole) und dem Konzept "shoe" (Schuh), "heel" (Absatz) und "shoe" usw. Die semantische Information (A), die im DM/MT **520** gehalten wird, identifiziert diese und andere semantische Beziehungen zwischen den Konzepten in der Domäne.

[0157] Wenn der Prozeß im MT-Analysierer **127** zum DM/MT **520** für semantische Information betreffend die Beziehung zwischen dem Konzept "valve" und dem Konzept "lever" geht, wird die Information im DM **137** den MT-Analysierer **127** nicht in die Lage versetzen, auszusagen, ob "lever" "is a part of" "valve" – die Kenntnis bzw. das Wissen über eine solche Beziehung ist einfach nicht vorhanden. Somit weiß der MT-Analysierer **127** immer noch nicht, ob die Phrase "with the lever" an das Wort "valve" angefügt werden sollte.

[0158] Wenn sich nun der MT-Analysierer **127** der Kontextinformation (B) zuwendet, findet er, daß das Verb "remove" drei Fälle annimmt: einen Nominativ (NOM), einen Akkusativ (ACC) und einen Instrumentellen (INS) (in einer tieferen Ebene der Analyse jedoch als die lateinische Grammatik unserer Schultage). Das bedeutet, daß "remove" in den folgenden Fallrahmen paßt.

$$\text{———————}_{\text{VERB}}(\text{NOM, ACC, INS})$$

[0159] Basierend auf diesem abstrakten Muster können wir Sätze wie folgt aufbauen.

NOM	VERB	ACC	INS
The workman	removed	the sand	with the shovel
Peter	has removed	the box	with the nail

usw.

[0160] Wenn das DM/MT Information über die Kombination mit der Präposition "with" und von Hauptwörtern mit dem semantischen Merkmal [+Instrument] enthält, bilden solche Kombinationen instrumentelle Phrasen. Diese Information ermöglicht dem Analysierer zu bestimmen, daß

a) da "lever" [+Instrument] ist, ist "with the lever" INS;

b) da "remove" den INS-Fall annehmen kann, ist die Phrase "with the lever" angebracht, paßt zusammen mit und wird als modifizierend für "remove" interpretiert.

[0161] Dennoch kann das DM **137** nur so reichhaltig sein, wie wir es aufbauen. In jenen Fällen, in denen die

semantische Information nicht so vollständig wie möglich entwickelt wurde, können die lexikalischen Eintragungen in der Domäne nicht in der Lage sein, den Zweideutigkeitsentfernungsprozeß zu unterstützen, der mittels dem MT-Analysierer **127** durchgeführt wird.

[0162] Man betrachte den Fall von "nail" (Nagel) in "Peter has removed the box with the nail" (Peter hat die Schachtel mit dem Nagel entfernt). Wenn das DM **137** die Information über Nägel als Teil eines hölzernen Rahmens enthält, jedoch nicht die Information enthält, daß "nails" (Nägel) [+Instrument] sind, dann kann der MT-Analysierer **127** unmöglich bestimmen, ob "with" sich mit "nail" zur Ausbildung einer instrumentellen Phrase kombiniert. Da der Analysierer nicht in der Lage ist, die strukturelle Zweideutigkeit aufzulösen, wird der Verfasser aufgefordert, sie aufzulösen. Wenn der durch den Verfasser eingegebene Text der Grammatiküberprüfung unterzogen wird, treten die folgenden Wechselwirkungen bzw. Interaktionen auf.
Eingabe des Verfassers, wenn überprüft: Peter has removed the box with the nail.
Nachricht der Grammatiküberprüfung **620**: Der Satz ist zweideutig.

1. Is the nail an instrument (Ist der Nagel ein Instrument)?
2. Hat die "box" (Schachtel) einen "nail" (Nagel)?

[0163] Sobald der Verfasser eine Interpretationsauswahl trifft, fügt die Überprüfung eine unsichtbare SGML-Kennzeichnung an den Satz, was dem System anzeigt, wie der Satz zu übersetzen ist.

[0164] Wie zuvor erwähnt, wird der MT-Analysierer **127** durch die Grammatiküberprüfung aufgerufen, um zu überprüfen, ob der Eingabetext oder ein IE (oder ein Teil davon) konform mit den grammatikalischen und semantischen Einschränkungen der CSL ist. Diesbezüglich erwidert ein bevorzugtes Ausführungsbeispiel mit einer strikten "Grünlicht, Rotlicht"-Nachricht für jeden Satz, wobei letzteres anzeigt, daß der Verfasser die Zusammensetzung der gekennzeichneten Sätze über die Verfassungsumgebung bzw. Verfasserumgebung korrigieren muß. Sobald der gesamte Eingabetext oder der IE als CSL-konform bzw. CSL-gemäß bestätigt wurde, kann er gespeichert oder für eine sofortige Übersetzung verschickt werden.

[0165] Bezugnehmend auf **Fig.** 8 ist ein Hochebenenfließdiagramm der Grammatiküberprüfung **620** (syntaktische Analyse) und der Zweideutigkeitsentfernungsüberprüfung **630** (semantische Analyse) gezeigt. Der Ausdruck bzw. das Wort "Satz" wird in der Folge verwendet, um sich auf eine Texteinheit zu beziehen, die die Überprüfung durch das Analysemodul **127** besteht oder daran scheitert. Die Einheit, die überprüft wird, kann tatsächlich eine nicht zu einem Satz gehörende Textkomponente sein, wie beispielsweise eine Überschrift, ein Titel oder Listenelemente, oder eine Unterschrift bzw. ein Untertitel oder anderer Text aus einer Graphik. Die Grammatiküberprüfung **620** erkennt Satzgrenzen und SGML-Elementgrenzen in einem SGML-gekennzeichneten Text. Sie überprüft jeden Satz, der nicht konform mit der CSL-Spezifikation ist. Dies umfaßt jeden Satz, bei dem nicht eine erfolgreiche Satzteilbestimmung durch das MT-Analysemodul **127** erreicht werden kann. Die Satzteilbestimmung kann aufgrund von Ursachen scheitern, die jene umfassen, die weiter unten aufgeführt sind, jedoch nicht auf diese eingeschränkt sind.

– der Satz umfaßt grammatikalische Konstruktionen, von denen das Analysemodul **127** keine Satzteilbestimmung durchführt. So ist es beispielsweise der Fall, wenn der Satz einen reduzierten Relativsatz enthält. Die Reduktion resultiert aus der Weglassung des Relativpronomens "that" und des Verbs "be" in einem Satz wie "Don't change the values that are programmed into the unit".

Eingabe des Verfassers, wenn überprüft: Don't change the values rogrammed into the unit.
Nachricht der Grammatiküberprüfung: Bei diesem Satz ist die Satzteilbestimmung schwierig.
Überprüfe auf eines der folgenden Probleme:

[0166] Dann fährt die Grammatiküberprüfung **620** damit fort, die typischen und häufigsten Situationen aufzulisten, wo eine Satzteilbestimmung erschwert wird, wenn nicht sogar unmöglich gemacht wird, durch die Verwendung von grammatikalischen Konstruktionen, die nicht im Repertoire der CSL enthalten sind.

– Die Satzzeichenverwendung im Satz ist nicht konform mit den CSL-Einschränkungen. Wie zuvor erwähnt, werden Satzzeichenmarkierungen und spezielle Zeichen, die nicht Teil der CSL in irgendeinem Kontext sind, durch die Vokabularüberprüfung **610** markiert bzw. gekennzeichnet. Jedoch führt die Vokabularüberprüfung **610** keine Satzteilbestimmung bezüglich der Eingabe durch, daher wird sie nicht Fälle berichten, in welchen solch ein Element in CSL existiert, jedoch im falschen Kontext verwendet wurde. Diese Fallart wird eine "Versagen"-Antwort von der Grammatiküberprüfung **620** auslösen.

– Ein CSL-Vokabularwort wurde in einer syntaktischen Form verwendet, die nicht für dieses Wort in CSL erkannt wird. Die Vokabularüberprüfung **610** wird einige dieser Fälle kennzeichnen; beispielsweise, wenn das Wort "test" in CSL umfaßt ist als ein Hauptwort jedoch nicht als ein Verb, wird die Vokabularüberprüfung berichten, daß die Vergangenheitsform "tested" nicht CSL ist. Jedoch wird die Vokabularüberprüfung **610**

die Verbgegenwartsform "tests" zulassen, da diese Form identisch zum Plural des CSL-Hauptworts "tests" ist. Dieser Fall wird eine "Versagen"- bzw. "Fehler"-Antwort von der Grammatiküberprüfung **620** auslösen.

[0167] Die Grammatiküberprüfung **620** verwendet das MT-Analysismodul bzw. - Analysemodul **127** (und das Domänenmodell **137**) zur Identifizierung von Sätzen, die nicht konform mit den CSL-grammatikalischen Einschränkungen sind, was als syntaktische Analyse bekannt ist, und im Block **805** gezeigt ist. Für jeden solchen Satz berichtet die Grammatiküberprüfung **620**, daß der Satz Nicht-CSL ist. Es ist auch möglich, daß ein Satz in CSL ist, jedoch zweideutig ist. Konsequenterweise sieht die vorliegende Erfindung eine semantische Analyse, wie in Block **710** gezeigt, vor. Wenn der Satz, der überprüft wird, nicht semantisch zweideutig ist, wird die Zweideutigkeitsentfernungsüberprüfung **630** eine Anzeige für die zwei oder mehr möglichen Bedeutungen dem Verfasser präsentieren und eine Klarstellung anfordern, wie in den Blöcken **815** und **825** gezeigt. In einem bevorzugten Ausführungsbeispiel, wenn ein Satz bei der Grammatiküberprüfung **620** und/oder der Zweideutigkeitsentfernungsüberprüfung **630** versagt bzw. durchfällt, hat der Verfasser die folgenden Optionen: Redigieren des Dokuments, in Fällen einer zweideutigen Lesweise, Zweideutigkeitsentfernung bezüglich des Satzes, nochmalige Überprüfung der gleichen Eingabe oder Fortfahren mit der Überprüfung ohne eine Redigierung bzw. Editierung.

[0168] Es sei bemerkt, daß die vorliegende Erfindung eine absolute Befolgung der Einschränkungen bezüglich des Vokabulars und der Grammatik ausführt, anstatt bloßer stilistischer Warnungen oder einfacher Fehlerdetektion (wie beispielsweise Subjekt-Verbübereinstimmung).

[0169] Wenn der Satz semantisch nicht zweideutig ist, dann wird er in die Zwischensprache bzw. Interlingua übersetzt, wie im Block **820** gezeigt. Sobald das Dokument die Grammatiküberprüfung **620** besteht, kann ein SGML-Kennzeichen, das die CSL-Genehmigung bezeichnet, in das Dokument eingefügt werden.

[0170] In einem bevorzugten Ausführungsbeispiel sieht die Grammatiküberprüfung **620** eine Bestehen/Versagen Rückkopplung für den Verfasser **160** vor. Jedoch kann eine spezifischere Rückkopplung als Bestehen/Versagen-Rückkopplung ausgeführt werden.

[0171] Für eine tiefergehende Diskussion der Grammatiküberprüfung, einschließlich der Zweideutigkeitsentfernung, siehe Tomita, M., "Sentence Disambiguation by Asking," Computers and Translation, 1: 39–51 (1986) und Carbonell, J. und M. Tomita, "Knowledge-Based Machine Translation, the CMU Approach," in S. Nirenburg (Hg.), Machine Translation: Theoretical and Methodological Issues, Cambridge: Cambridge University Press, Seiten 68–89 (1987).

E. Maschinenübersetzung

[0172] Der MT **120** (Machine Translation = Maschinenübersetzer) ist ein zwischensprachenartiges Maschinenübersetzungssystem. In solchen Systemen kommen die eingeschränkte Quellensprache (CSL) und die Zielsprache niemals in direkten Kontakt. Die Verarbeitung in solchen Systemen geschieht im allgemeinen in zwei Stufen. Zuerst wird die Bedeutung des CSL-Textes in einer sprachunabhängigen formalen Sprache repräsentiert, Interlingua bzw. die Zwischensprache genannt, und zweitens diese Bedeutung unter Verwendung der lexikalischen Einheiten und syntaktischen Konstruktionen der Zielsprache ausgedrückt.

[0173] Interlingua-MT-Systeme sind wie andere Typen der MT-Systeme im Stand der Technik gut bekannt. Detaillierte Beschreibungen von diesen verschiedenen Herangehensweisen an eine Maschinenübersetzung können gefunden werden in Hutchins, Machine Translation: Past, Present, Future, Ellis Horwood, Ltd.; Chichester, UK, 1986, und Zarechnak, The History of Machine Translation, in Henisz-Dostert, McDonald, Zarechnak, Hg., Machine Translation. Trends in Linguistics: Studies and Monographs, Vol. 11, The Hague, Mouton, 1979.

[0174] Die Bedeutung des CSL-Textes **350** wird in dem speziell konstruierten Wissensrepräsentationsschema repräsentiert, das Interlingua genannt wird (was im Stand der Technik gut bekannt ist). Interlingua wiederum wird in einer Rahmennotation repräsentiert und kann somit als eine Art von semantischem Netzwerk betrachtet werden. Wie andere künstliche oder formale Sprachen hat Interlingua sein eigenes Lexikon und seine eigene Syntax. Das Lexikon basiert auf der Domäne, aus der die übersetzten Texte genommen werden (beispielsweise Computerwartung, Raumerforschung, usw.). Somit sind Interlingua-"Hauptwörter" "Objektkonzepte" in der Ontologie; Interlingua-Verben entsprechen grob den "Ereignissen" in der Ontologie; und Interlingua-Adjektive und Adverbien sind die verschiedenen "Eigenschaften", die in der Ontologie definiert sind. Die Ontologie bildet ein dicht verbundenes Netzwerk für die verschiedenen Typen der Konzepte, was das Domä-

nenmodell genannt wird.

[0175] Bezugnehmend auf **Fig.** 3 und **Fig.** 9 enthält die Maschinenübersetzungs-(MT)-Komponente **120** des IATS **105** zwei Hauptabschnitte. Der erste, der CSL-Analysierer **127**, führt die erste Verarbeitungsstufe des Repräsentierens des CSL-Textes in Interlingua durch. Der zweite Hauptabschnitt, der Zielsprachengenerator **123**, übersetzt die Interlinguarepräsentation der "CSL-genehmigten" Texte in eine Zielsprache (beispielsweise Französisch, Japanisch, Spanisch). Bei der Durchführung dieser Aufgaben läuft die MT-Komponente **120** als ein oder mehrere unabhängige Servermodule, die Übersetzungsanforderungen von einem menschlichen Übersetzungscontroller (nicht gezeigt) annehmen bzw. akzeptieren. Während der Zielsprachengenerierung bzw. Zielsprachenerzeugung bildet der Zielsprachengenerator **123** den Interlinguatext **260** in die geeigneten Einheiten der Zielsprachensyntax ab, um einen hochqualitativen Ausgabetext **950** zu erzeugen, der keine Nacheditierung erfordert.

[0176] Sobald das MT-Analysemodul **127** einen Interlinguatext **260** für einen zertifizierten bzw. genehmigten CSL-konformen IE produziert hat, kann diese Interlingua weggespeichert werden, geliefert oder sofort in einen Zielsprachen-IE konvertiert werden, oder in einen IE in einer jeden der vielen Zielsprachen, und zwar durch den Generator **123** (der einen Semantik-zu-Syntax-Abbilder umfaßt und ein Erzeugungswerkzeug (Generation Kit) (Tomita M. und E. Nyberg, The Generation Kit and Transformation Version 3.2 User's Manual, Technical Memo (1988), erhältlich vom Center for Machine Translation, Carnegie Mellon University, Pittsburgh, Pa.). Der MT-Analysierer **127** und der MT-Generator **123** wechselwirken auf zwei Wegen. Erstens wird die Ausgabe des Ersteren zur Eingabe des Letzteren, und zweitens teilen sie gemeinsam einige externen Wissensquellen, insbesondere das Domänenmodell **137**.

[0177] Das MT-System **120** ist unterteilt, wie in **Fig.** 9 gezeigt. Die Analyse besteht aus einem Satzteilbestimmer **910** und einem Interpretierer bzw. Dolmetscher **920**. Die andere Hälfte des MT **120** kann in einen Abbilder **930** und einen Generator **940** unterteilt werden. Die ovalen Kreise in **Fig.** 9 stehen für die Daten, die von den zwei Hauptsoftwaremodulen erzeugt werden und zwischen ihnen weitergegeben werden.

[0178] Das DM **137** (und insbesondere das MT/DM **520**) wird auf drei verschiedene Arten während der Übersetzung verwendet: (1) der Satzteilbestimmer **910** verwendet das DM **137**, um mögliche Anfügungen einzuschränken (unter Verwendung einer strikten Teilkategorisierung der Argumente und unter Verwendung von Modifizierern während der syntaktischen Satzteilbestimmung); (2) der Interpretierer **920** verwendet das DM **137**, um die geeigneten Domänenkonzepte während der Interpretierung beispielhaft darzustellen; (3) der Abbilder **930** verwendet das DM **137**, um die geeignete Zielrealisierung für jedes Interlinguakonzept auszuwählen.

[0179] Der MT **120** läuft als einer oder mehrere Serverprozesse bzw. Dienstprozesse. Jeder solcher MT-Prozeß akzeptiert Übersetzungsanforderungen vom FMS **110** und gibt die Ergebnisse zurück. Die Anforderungen enthalten SGML-gekennzeichneten CSL-Text und die Ergebnisse enthalten SGML-gekennzeichnete Zielsprachenübersetzungen. Da Übersetzungen in mehr als eine Sprache gleichzeitig stattfinden können, umfassen die Anfragen ebenso die gewünschte Zielsprache. Da die MT-Serverprozesse durch die Zielsprache spezialisiert sind, wird eine Leit- bzw. Rootingfunktion miteinbezogen. Diese Rootingfunktion wird automatisch durch das FMS **110** durchgeführt. Der präzise Satz der MT-Prozesse, die bei einer gegebenen Zeit ablaufen, und ihre Verteilung über die Maschinen wird durch das FMS **110** festgelegt, welches die Mischung gemäß dem Satz der Übersetzungsaufgaben bzw. Übersetzungsjobs, die während einer besonderen Zeit anstehen, modifizieren wird.

[0180] Bezugnehmend auf **Fig.** 9, besteht der CSL-Analysierer **127** aus zwei miteinander verbundenen Komponenten – einen syntaktischen Satzteilbestimmer **910** und einen semantischen Interpretierer **920**. Der semantische Interpretierer **920** ist ebenso im Stand der Technik als ein "Abbildungsregelinterpretierer" ("mapping rule interpreter") bekannt. Der syntaktische Satzteilbestimmer **910** erhält die CSL-Text-305-Eingabe und erzeugt eine syntaktische Struktur für dieselbe. Der syntaktische Satzteilbestimmer **910** verwendet eine LFG-artige Grammatik. Lexikalische Funktionalgrammatik (LFG) ist eine formalisierte Grammatik, die im Stand der Technik für Maschinenübersetzungen gut bekannt ist. Als ein Ergebnis ist die resultierende syntaktische Struktur eine LFG f-Struktur **960**. Sobald die f-Struktur für den CSL-Satz **960** erzeugt ist, startet der semantische Interpretierer **920** mit der Anwendung der Abbildungsregeln, um die lexikalischen Einheiten der Quellensprache und syntaktischen Konstruktionen gegen ihre Zwischensprachenübersetzungen zu ersetzen. Lexikalische Einheiten werden in Ereignisse bzw. Beispiele der Domänenkonzepte abgebildet (beispielsweise das Wort "data" wird in Zwischensprache "Information" abgebildet), während die syntaktischen Strukturen in konzeptuelle Beziehungen abgebildet werden (beispielsweise werden Subjekte von Sätzen oft in die "handelnden"- bzw.

"Agent"-Beziehungen in der Zwischensprache abgebildet). Siehe Mitamura, The Hierarchical Organization of Predicate Frames for Interpretive Mapping in Natural Language Processing, Center for Machine Translation, Carnegie Mellon University (Mai 1990).

[0181] Der MT-Analysierer **127**, und zwar geführt durch das Analysewissen (Datendateien), übersetzt einen Eingabesatz des CSL-Textes **305** in der Quellensprache in eine semantische Rahmenrepräsentation der Bedeutung dieses Satzes. Die Wissensstrukturen, die in der Analysephase zum Tragen kommen, sind die Analysegrammatiken, die Abbildungsregeln und das Konzeptlexikon.

[0182] Ein erster Teil der Analyse ist der Satzteilbestimmungsprozeß, der durch die syntaktische Analyse des Eingabesatzes angetrieben wird. Der Satzteilbestimmer **910** verwendet die semantischen Einschränkungen, die im Konzeptlexikon ausgeführt sind (Domänenmodell), um seine Behandlung von syntaktischen Zweideutigkeiten zu leiten, die während seiner Analyse der Eingabe angetroffen werden. Die Abbildungsregeln vermitteln zwischen den syntaktischen Analysegrammatiken und dem Konzeptlexikon.

[0183] Die Ausgabe dieser Analyse sind syntaktische f-Strukturen, die die ganze anwendbare semantische Information enthalten. Diese Struktur kann weiter durch den zweiten Teil des MT-Analysierers **127** verarbeitet werden, um eine semantisch organisierte Rahmenrepräsentation zu erzeugen, und zwar in der Form der beispielhaften Darstellung der relevanten Konzepte aus dem Konzeptlexikon, die während der Satzteilbestimmung des Satzes angetroffen wurden. Der MT-Analysierer **127** gelangt zu dieser Form durch Wiedergewinnung bzw. Herausholen der semantischen Merkmale der f-Struktur; diese Merkmale enthalten alle relevante semantische Information.

[0184] Der syntaktische Satzteilbestimmer **910**, der in der vorliegenden Erfindung verwendet wird, ist im Stand der Technik gut bekannt und detailliert in Tomita und Carbonell, The Universal Parser Architecture for Knowledge-Based Machine Translation, Technical Report, Center for Machine Translation, Carnegie Mellon University (Mai 1987), und Tomita (Hg.) u. a., The Generalized LR Parser/Compiler Version 8.1: User's Guide, Technical Memo, Center for Machine Translation, Carnegie Mellon University (April 1988) beschrieben.

[0185] Einer der Vorteile der Zwischensprachenübersetzungssysteme gegenüber anderen Typen der MT-Systeme ist der, daß die Interlingua bzw. Zwischensprache **260** sprachunabhängig ist; das bedeutet, daß die Subjekt- und die Zielsprachen niemals in direktem Kontakt sind. Dies erlaubt die Konstruktion von einem Maschinenübersetzungssystem, bei welchem potentiell jegliche Quellen- und Zielsprachen ausgewählt werden könnten, während minimale Modifikationen bezüglich der Rechnerstruktur bzw. Computerstruktur erforderlich sind. Klarerweise muß dann jegliches solches System in der Lage sein, bei einer Anzahl von Quellensprachen die Satzteilbestimmung durchzuführen. Demgemäß wird ein universeller Satzteilbestimmer benötigt, der eine Sprachgrammatik als eine Eingabe annimmt, anstatt die Grammatik in den Interpretierer bzw. Dolmetscher selbst einzubauen. Dies erlaubt eine größere Erweiterbarkeit und Verallgemeinerung.

[0186] In anderen Worten, wenn mehrere Sprachen behandelt werden, ist die linguistische Struktur nicht länger eine universelle Invariante, die sich über alle Anwendungen überträgt, (wie es für reine englischsprachige Satzteilbestimmer der Fall war), sondern statt dessen ist sie eine weitere Dimension der Parameterbildung und Erweiterbarkeit. Jedoch kann semantische Information über die Sprachen invariant bleiben (natürlich jedoch nicht über Bereiche bzw. Domänen). Demgemäß ist es ein Kernpunkt, die semantischen Wissensquellen getrennt von den syntaktischen zu halten, so daß, wenn neue linguistische Information zugefügt wird, sie über alle semantischen Domänen angewendet wird, und wenn neue semantische Information zugefügt wird, sie über alle relevanten Sprachen angewendet wird. Der universale Satzteilbestimmer versucht diese Faktorisierung zu bewerkstelligen, ohne größere Konzessionen entweder bezüglich der Laufzeiteffizienz oder der semantischen Genauigkeit machen zu müssen.

[0187] Der Satzteilbestimmer **910** ist durch drei Arten von Wissensquellen charakterisiert. Eine enthält syntaktische Grammatiken für verschiedene Sprachen, eine weitere enthält semantische Wissensbasen für verschiedene Domänen und die dritte enthält Sätze von Regeln, die syntaktische Formen (Wörter und Phrasen) in die semantische Wissensstruktur abbilden. Jede der syntaktischen Grammatiken ist völlig unabhängig von jeglicher spezifischer Domäne; gleichfalls ist die semantische Wissensbasis unabhängig von jeglicher spezifischer Domäne; gleichfalls ist die semantische Wissensbasis unabhängig von jeglicher spezifischer Sprache.

[0188] Ferner sind die Abbildungsregeln sowohl sprach- als auch domänenabhängig, und verschiedene Sätze von Abbildungsregeln werden für jede Sprach/Domänenkombination erzeugt. Syntaktische Grammatiken, Domänenwissensbasen und Abbildungsregeln werden in einer hochabstrakten, durch den Menschen lesbaren

Weise geschrieben. Diese Organisation macht sie leicht für eine Erweiterung oder Modifizierung, jedoch möglicherweise maschinenineffizient für einen Laufzeitsatzteilbestimmer.

[0189] Die Aufgabe bzw. Funktion des Abbildungsregelinterpretierers **920** ist es, die syntaktischen und semantischen Strukturen einer Satzteilbestimmung zu erzeugen und zu manipulieren und darüber hinaus diese Strukturen simultan zu erzeugen.

[0190] Der universale bzw. universelle Satzteilbestimmer **910** erzeugt alle möglichen, d. h. gültigen f-Strukturen, die aus dem satzteilbestimmten Sätzen abgeleitet werden können. Jede dieser syntaktischen f-Strukturen besitzt semantische Merkmale, und zwar werden gemäß der LFG-Theorie diese Merkmale zur selben Zeit erzeugt, wie der Rest der syntaktischen f-Struktur. Die semantische Komponente kann somit als ein zusätzliches Merkmal der f-Strukturen betrachtet werden.

[0191] Somit ist die semantische Komponente ein "sichtbarer" Teil der syntaktischen Satzteilbestimmung. Die Herangehensweise der simultanen Erzeugung der syntaktischen und semantischen Strukturen hat ein System erzeugt, das in der Lage ist, "bedeutungslose" Teilsatzteilbestimmungen vor deren Vollendung zu eliminieren. Semantiken werden der syntaktischen Struktur hinzugefügt, wenn auf das Lexikon für die Definition eines Wortes zugegriffen wird. Ein weiterer Teil der Definition eines Wortes ist ein Satz von strukturellen Abbildungsregeln. Diese Abbildungsregeln werden verwendet, wenn syntaktische Gleichungen in Grammatikregeln Information zu einer syntaktischen Struktur hinzufügen.

[0192] Die Zielsprachengeneratorkomponente **123** nimmt den Zwischensprachentext **260** als ihre Eingabe und erzeugt einen Zielsprachentext **950** als ihre Ausgabe. Der Zielsprachengenerator **123** besteht aus zwei Hauptmodulen, einem semantischen und einem syntaktischen. Das semantische führt die Funktion der lexikalischen Auswahl für die Zielsprache und der Auswahl der syntaktischen Konstruktionen der Zielsprache durch; es wird bei diesen Aufgaben durch ein Erzeugungslexikon bzw. Abbildungsregeln für die Erzeugungsstruktur unterstützt. Die Ausgabe dieses Moduls ist eine f-Struktur des Zielsprachsatzes, der durch das System ausgegeben wird.

[0193] Das Ziel des Erzeugungsmoduls ist es, Zielsprachsätze aus den Rahmen des Interlinguatextes **260** zu erzeugen, die durch den CSL-Analysierer **127** erzeugt werden. Es gibt drei Hauptschritte bei der Erzeugung:

1. Lexikalische Auswahl.
   Für jedes Konzept in der Zwischensprache muß der geeignetste lexikalische Punkt ausgewählt werden.
2. F-Strukturerzeugung.
   Eine syntaktische Funktionalstruktur, die die grammatikalische Struktur der Zieläußerung bestimmt, muß aus den Zwischensprachentextrahmen erzeugt werden.
3. Syntaktische Erzeugung.
   Die syntaktische Funktionalstruktur wird durch die Erzeugungsgrammatik verarbeitet, um einen Zielsprachsatz zu erzeugen.

[0194] Die Konstruktion des Erzeugungsmoduls **940** kombiniert jüngste Forschung auf dem Gebiet der lexikalischen Auswahl mit einem Abbildungs- und - Erzeugungsparadigmus, der in früheren bzw. vorangehenden Übersetzungssystemen verwendet wurde.

[0195] Für eine tiefergehende Diskussion von Maschinenübersetzung und dem spezifischen Design bzw. der Konstruktion und dem Betrieb der zuvor beschriebenen Module siehe Nirenburg u. a., Machine Translation: A Knowledge-Based Approach, Morgan Kaufmann Publishers, Inc. (1992), Sommers & Hutchins, Introduction to Machine Translation, Academic Press, London (Oktober 1991), Mitamura u. a., An Efficient Interlingua Translation System for Multi-lingual Document Production, Proceedings of Machine Translation Summit III, Washington D.C. (Juli 2–4, 1991), Nirenburg, S., "World Knowledge and Text Meaning", in K. Goodman und S. Nirenburg (Hg.), The KBMT Project: A Case Study in Knowledge-Based Machine Translation, San Mateo, Kalif.: Morgan Kaufmann, KBMT-89 Project Report, erhältlich vom Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA (Telefonnummer (412) 268-6591) (Vierte Auflage: März 1990), S. Nirenburg (Hg.), Machine Translation: Theoretical and Methodological Issues, Cambridge: Cambridge University Press, Seiten 68–89 (1987), und Carbonell u. a., Steps Toward Knowledge-Based Machine Translation, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. PAMI-3, Nr. 4 (Juli 1981).

[0196] Während die Erfindung insbesondere unter Bezugnahme auf bevorzugte Ausführungsbeispiele davon gezeigt und beschrieben wurde, sei durch den Fachmann verstanden, daß verschiedene Veränderungen bezüglich der Form und bezüglich von Einzelheiten daran durchgeführt werden können, ohne vom Geist und Um-

fang der Erfindung abzuweichen.

Patentansprüche

1. Computerbasierendes System (**105**) für einsprachige Dokumententwicklung, das folgendes aufweist:
einen Texteditor (**140**) angepaßt zur interaktiven Aufnahme eines Eingabetextes eines Verfassers (**160**) in einer Quellensprache; und
einen Spracheditor (**130**), der eine Erweiterung des Texteditors (**140**) ist, der interaktiv lexikalische Einschränkungen und grammatikalische Einschränkungen auf einen durch den Verfasser (**160**) verwendeten Teilsatz der natürlichen Sprache erzwingt, um den Eingabetext zu erzeugen, wobei der Verfasser interaktiv beim Erzwingen der lexikalischen Einschränkungen und der grammatikalischen Einschränkungen bezüglich des Eingabetextes unterstützt wird, um so einen unzweideutigen, eingeschränkten Text zu erzeugen;
ein Maschinenübersetzungssystem (**105**) ansprechend auf den Spracheditor (**130**), das zur Übersetzung des unzweideutigen, eingeschränkten Textes in eine Fremdsprache konfiguriert ist;
gekennzeichnet durch
ein Domainenmodell (**137**), welches mit dem Spracheditor (**130**) kommuniziert, wobei das Domainenmodell (**137**) ein vorbestimmtes Domainenwissen und ein linguistisches, semantisches Wissen über die lexikalischen Einheiten und deren Kombinationen vorsieht, um so den Spracheditor (**130**) beim Erzwingen der lexikalischen und grammatikalischen Einschränkungen zu unterstützen, wobei das Domainenmodell (**137**) ein dreigeteiltes Domainenmodell ist, wobei das dreigeteilte Domainenmodell folgendes aufweist:
einen Kern (**510**), der lexikalische Information enthält, die durch den Spracheditor (**130**) und das Maschinenübersetzungssystem (**105**) angefordert wird, wobei die lexikalische Information lexikalische Eintragungen bzw. Punkte innerhalb des Teilsatzes der natürlichen Sprache zusammen mit assoziierten semantischen Konzepten, Wortklassen und morphologische Information aufweist,
ein Spracheditordomainenmodell (**530**), welches Information enthält, die nur durch den Spracheditor (**130**) angefordert wird, wobei die Information zumindest eines der folgenden umfaßt: einen Teilsatz der natürlichen Sprachen aus Synonymen für Eintragungen bzw. Punkte, die nicht innerhalb des Teilsatzes der natürlichen Sprache sind, ein Wörterbuch für die Definitionen der lexikalischen Eintragungen und einen Beispielsatz für die Verwendung der lexikalischen Eintragungen, und
ein Maschinenübersetzungsdomainenmodell (**520**), welches Information enthält, die nur vom Maschinenübersetzungsdomainenmodell (**520**) angefordert wird, und zwar einschließlich einer Hierarchie von Konzepten, die für die unzweideutige Abbildung und eine semantische Verifizierung der Übersetzung verwendet wird.

2. Computerbasierendes System (**105**) gemäß Anspruch 1, wobei der Spracheditor (**130**) eine Grammatiküberprüfung (**620**) aufweist, die Mittel (**630**) für eine interaktive Zweideutigkeitsvermeidung aufweist.

3. Computerbasierendes System (**105**) gemäß Anspruch 1, wobei der Spracheditor (**130**) eine Vokabelüberprüfung (**610**) zur Überprüfung des Eingabetextes gegen ein zugelassenes Lexikon und für den Vorschlag von Alternativen aufweist.

4. Computerbasierendes Verfahren zur einsprachigen Dokumententwicklung, das die folgenden Schritte aufweist:
(1) Eingabe eines Eingabetextes in einer Quellensprache in einen Texteditor (**140**);
(2) Überprüfung mittels eines Spracheditors (**130**) des Eingabetextes gegen einen vorbestimmten Satz von Einschränkungen;
(3) Liefern eines interaktiven Feedbacks bzw. einer Rückkopplung an den Verfasser (**160**) bezüglich des Eingabetextes, wobei die interaktive Rückkopplung anzeigt, ob der vorbestimmte Satz von Einschränkungen eingehalten wird;
(4) nach Vollendung des Schrittes (3) Erzeugen eines unzweideutigen, eingeschränkten Textes;
(5) Übersetzung mittels eines Maschinenübersetzungssystems (**105**) des unzweideutigen eingeschränkten Textes in eine Fremdsprache; wobei das Verfahren gekennzeichnet ist dadurch, daß
der vorbestimmte Satz von Einschränkungen in einem Domainenmodell (**137**) gespeichert wird, das ein vorbestimmtes Domainenwissen und ein linguistisches Semantikwissen über die lexikalischen Einheiten und ihrer Kombinationen liefert, wobei der vorbestimmte Satz von Einschränkungen einen Satz von Quellenteilsprachregeln umfaßt, die das Vokabular und die Grammatik betreffen, wobei das Domainenmodell (**137**) ein dreigeteiltes Domainenmodell ist, wobei das dreigeteilte Domainenmodell folgendes aufweist:
einen Kern (**510**), der lexikalische Information enthält, die durch den Spracheditor (**130**) und das Maschinenübersetzungssystem (**105**) angefordert wird, wobei die lexikalische Information folgendes aufweist: lexikalische Eintragungen bzw. Punkte, die den vorbestimmten Satz von Einschränkungen erfüllen, zusammen mit assoziierten semantischen Konzepten, Wortklassen und morphologischer Information,

ein Spracheditordomainenmodell (**530**), welches Information enthält, die nur durch den Spracheditor (**130**) angefordert wird, wobei diese Information zumindest eines der folgenden umfaßt: einen Teilsatz von Synonymen für die Eintragungen, die nicht dem vorbestimmten Satz von Einschränkungen genügen, eine Wörterbuchdefinition der lexikalischen Eintragungen und einen Beispielsatzes für die Verwendung der lexikalischen Eintragungen, und
ein Maschinenübersetzungsdomainenmodell (**520**), welches Information enthält, die nur vom Maschinenübersetzungsdomainensystem (**105**) angefordert wird, wobei das Maschinenübersetzungsdomainenmodell (**520**) eine Hierarchie von Konzepten aufweist, die für die unzweideutige Abbildung und die semantische Verifizierung der Übersetzung verwendet wird;
wobei die interaktive Rückkopplung nachfolgend zur Bezugnahme auf das Domainenmodell (**137**) durchgeführt wird, was das notwendige Domainenwissen und das linguistische Semantikwissen über die lexikalischen Einheiten und ihre Kombinationen vorsieht, und ferner die Grammatik eines Teilsatzes einer natürlichen Sprache.

5. Computerbasierendes Verfahren gemäß Anspruch 4, wobei der vorbestimmte Satz von Einschränkungen einen Satz von Quellenteilsprachregeln betreffend das Vokabular und Grammatik aufweist, wobei die interaktive Rückkopplung durchgeführt wird, um den Eingabetext dem Satz der Quellenteilsprachregeln anzupassen und Zweideutigkeiten zu eliminieren.

6. Computerbasierendes Verfahren gemäß Anspruch 4, wobei Schritt 4 folgendes aufweist:
Überprüfung auf syntaktische grammatikalische Fehler und semantische Zweideutigkeiten im eingeschränkten Qellentext durch Konsultieren des Domainenmodells (**137**); und
Vorsehen der interaktiven Rückkopplung an den Verfasser (**160**), um syntaktische grammatikalische Fehler und semantische Zweideutigkeiten im eingeschränkten Quellentext zu entfernen, und zwar zur Erzeugung eines unzweideutigen eingeschränkten Textes.

7. Computerbasierendes System (**105**) nach Anspruch 1, das weiter folgendes aufweist:
Mittel zur Markierung mit einem Kennzeichen eines Teils des Eingabetextes, welcher in einen unzweideutigen, eingeschränkten Text durch die interaktive Forcierung bzw. Erzwingung umgewandelt wurde, wobei das Kennzeichen eine linguistische Charakteristik des Teils des Eingabetextes anzeigt.

8. System (**105**) nach Anspruch 7, das weiter Mittel zur Markierung mit einem Kennzeichen eines Teils des Eingabetextes aufweist, der durch die interaktive Erzwingung in einen unzweideutigen, eingeschränkten Text umgewandelt wurde, wobei das Kennzeichen eine Übersetzbarkeit anzeigt.

9. System (**105**) nach Anspruch 7, wobei das Maschinenübersetzungsystem (**123**) in einen Übersetzungsserverumfeld betrieben wird, das mehreren Verfassern (**160**) die Verwendung des Systems erlaubt.

10. System (**105**) nach Anspruch 7, wobei der Verfasser (**160**) eine Workstation bzw. Arbeitsstation bedient bzw. betreibt, die ein Teil eines Computernetzwerkes ist.

11. System (**105**) nach Anspruch 7, wobei das Maschinenübersetzungssystem (**123**) einen Übersetzer (**920**) aufweist, der zur Übersetzung des unzweideutigen eingeschränkten Quellentextes in Interlingua bzw. eine Zwischensprache konfiguriert ist.

12. System (**105**) nach Anspruch 7, wobei der Spracheditor (**130**) eine Interaktion mit dem Verfasser (**160**) mittels Stapelverarbeitungsbetriebsweise vorsieht.

13. System (**105**) nach Anspruch 7, das weiter einen Graphikeditor (**150**) aufweist, der für das Erzeugen von Textkennungen adaptiert ist, wobei die Textkennungen durch den Verfasser (**160**) editiert, und zwar mit der Hilfe des Spracheditors (**130**), und nachfolgend durch das Maschinenübersetzungssystems (**123**) übersetzt werden können.

14. System (**105**) nach Anspruch 7, wobei die eingeschränkte Sprache ein Teilsatz einer natürlichen Sprache ist, wobei die eingeschränkte Sprache lexikalisch und grammatisch spezifiziert ist.

15. System (**105**) nach Anspruch 7, wobei der Spracheditor (**130**) eine Vokabelüberprüfung (**610**) und eine Grammatiküberprüfung (**620**) aufweist.

16. System (**105**) nach Anspruch 15, wobei die Vokabelüberprüfung (**610**) den Eingabetext gegen ein zu-

gelassenes Lexikon überprüft und Alternativen für nicht-Lexikonwortwahlen vorschlägt.

17. System (**105**) nach Anspruch 15, wobei die Grammatiküberprüfung (**620**) die Einhaltung der vorher bestimmten grammatikalischen Regeln überprüft und Alternativen für undefinierte grammatikalische Strukturen vorschlägt.

18. System (**105**) nach Anspruch 15, wobei die Grammatiküberprüfung (**620**) eine Rückkopplung an den Verfasser (**160**) liefert, und zwar lexikalische Zweideutigkeiten und strukturelle Zweideutigkeiten betreffend.

19. System (**105**) nach Anspruch 15, wobei die Grammatiküberprüfung (**620**) Mittel zur interaktiven Zweideutigkeitsentfernung vorsieht.

20. System (**105**) nach Anspruch 15, wobei die Vokabelüberprüfung (**610**) eine Buchstabierüberprüfung (**615**) aufweist.

21. System (**105**) nach Anspruch 15, wobei die Vokabelüberprüfung (**610**) zur Identifizierung von Worten konfiguriert ist, die nicht in der eingeschränkten Quellensprache enthalten sind.

22. System (**105**) nach Anspruch 7, wobei der Eingabetext in Blöcken von Informationselementen vorgesehen ist.

23. System (**105**) nach Anspruch 22, wobei die Informationselemente Kennzeichen enthalten, die es ermöglichen, daß die Informationselemente (**410**) bezüglich ihres Inhalts und ihrer logischen Struktur beschrieben werden.

24. System (**105**) nach Anspruch 7, das weiter Speichermittel zur Speicherung des unzweideutigen, eingeschränkten Textes für eine spätere Verwendung aufweist.

25. System nach Anspruch 7, wobei das Kennzeichen anzeigend für den Inhalt und die logische Struktur ist.

26. System nach Anspruch 7, wobei das Kennzeichen anzeigend für eine definierte Bedeutung des Teils ist, der vom Verfasser ausgewählt wurde.

Es folgen 10 Blatt Zeichnungen

Anhängende Zeichnungen

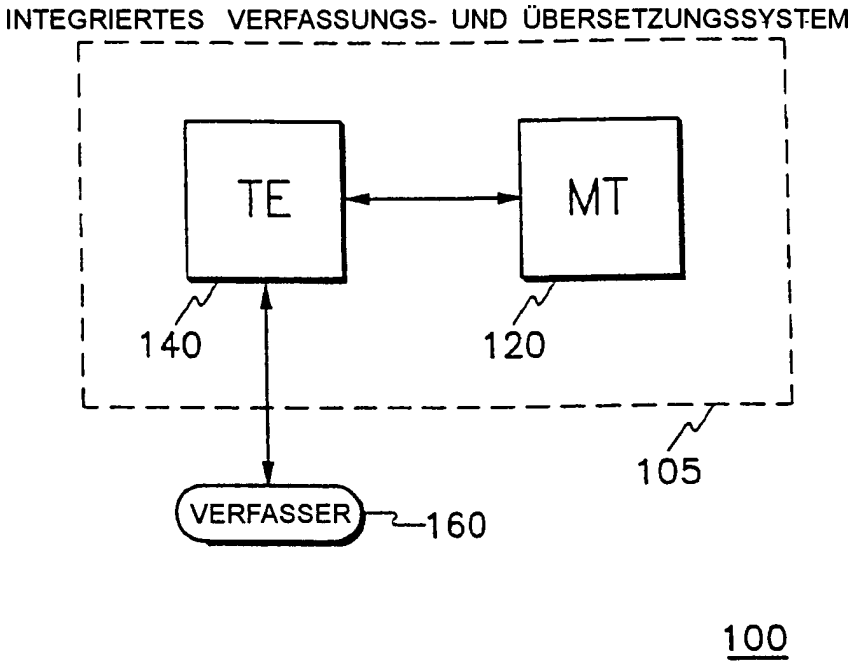

Fig. 1A

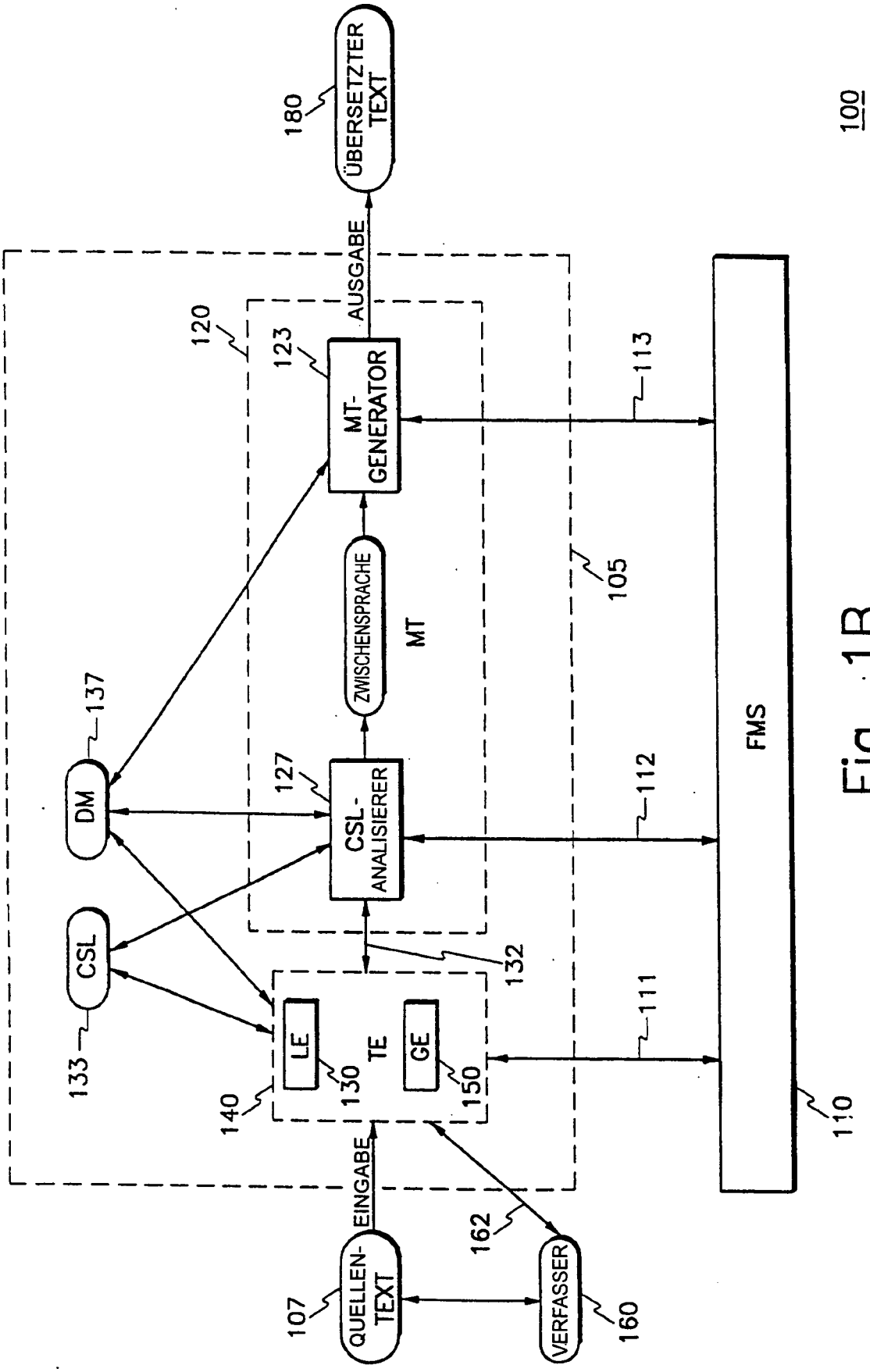
100
180
ÜBERSETZTER TEXT
AUSGABE
120
123
MT-GENERATOR
113
105
ZWISCHENSPRACHE
MT
137
DM
127
CSL-ANALISIERER
112
FMS
CSL
132
133
LE
130
TE
GE
150
111
140
110
EINGABE
162
107
QUELLEN-TEXT
VERFASSER
160

Fig. 1B

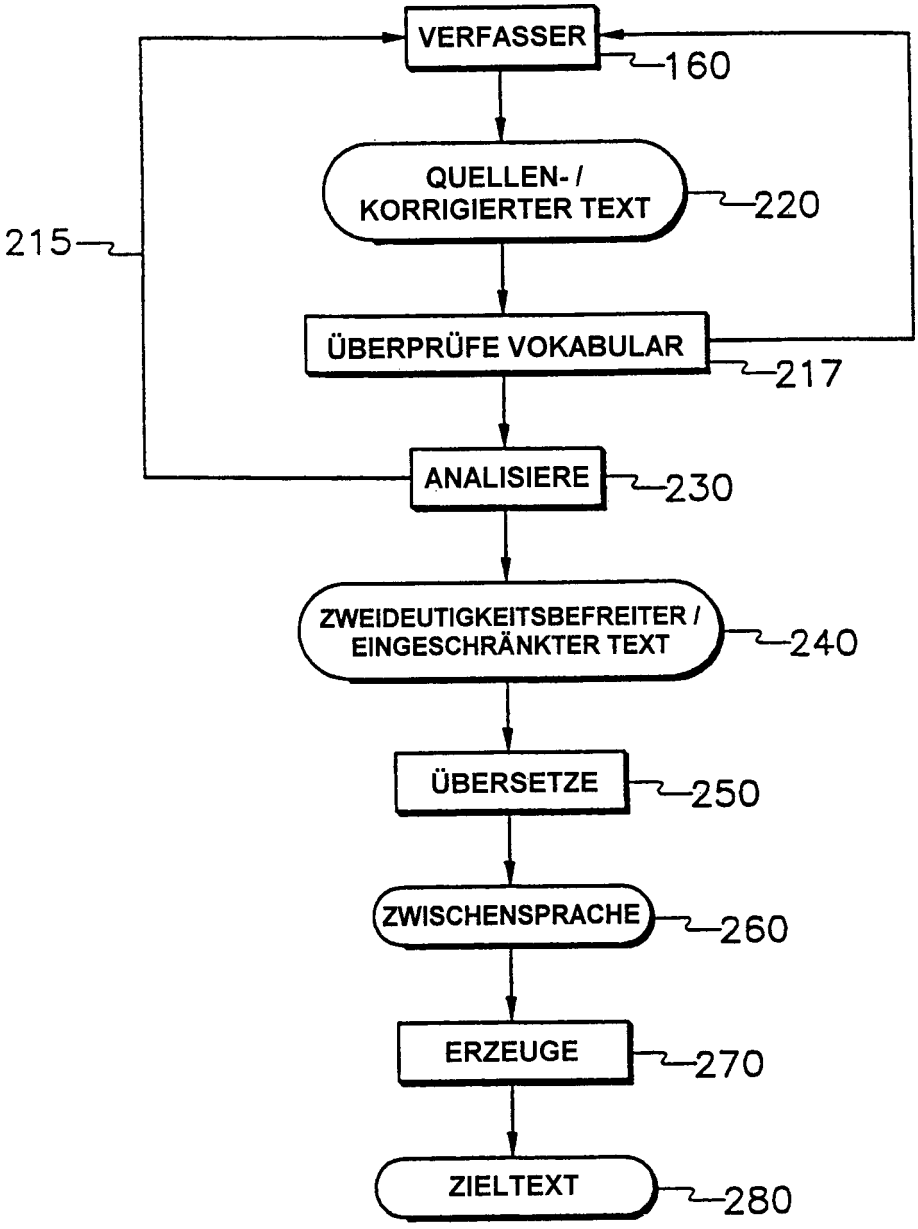
VERFASSER
160
QUELLEN- / KORRIGIERTER TEXT
220
215
ÜBERPRÜFE VOKABULAR
217
ANALISIERE
230
ZWEIDEUTIGKEITSBEFREITER / EINGESCHRÄNKTER TEXT
240
ÜBERSETZE
250
ZWISCHENSPRACHE
260
ERZEUGE
270
ZIELTEXT
280

Fig. 2

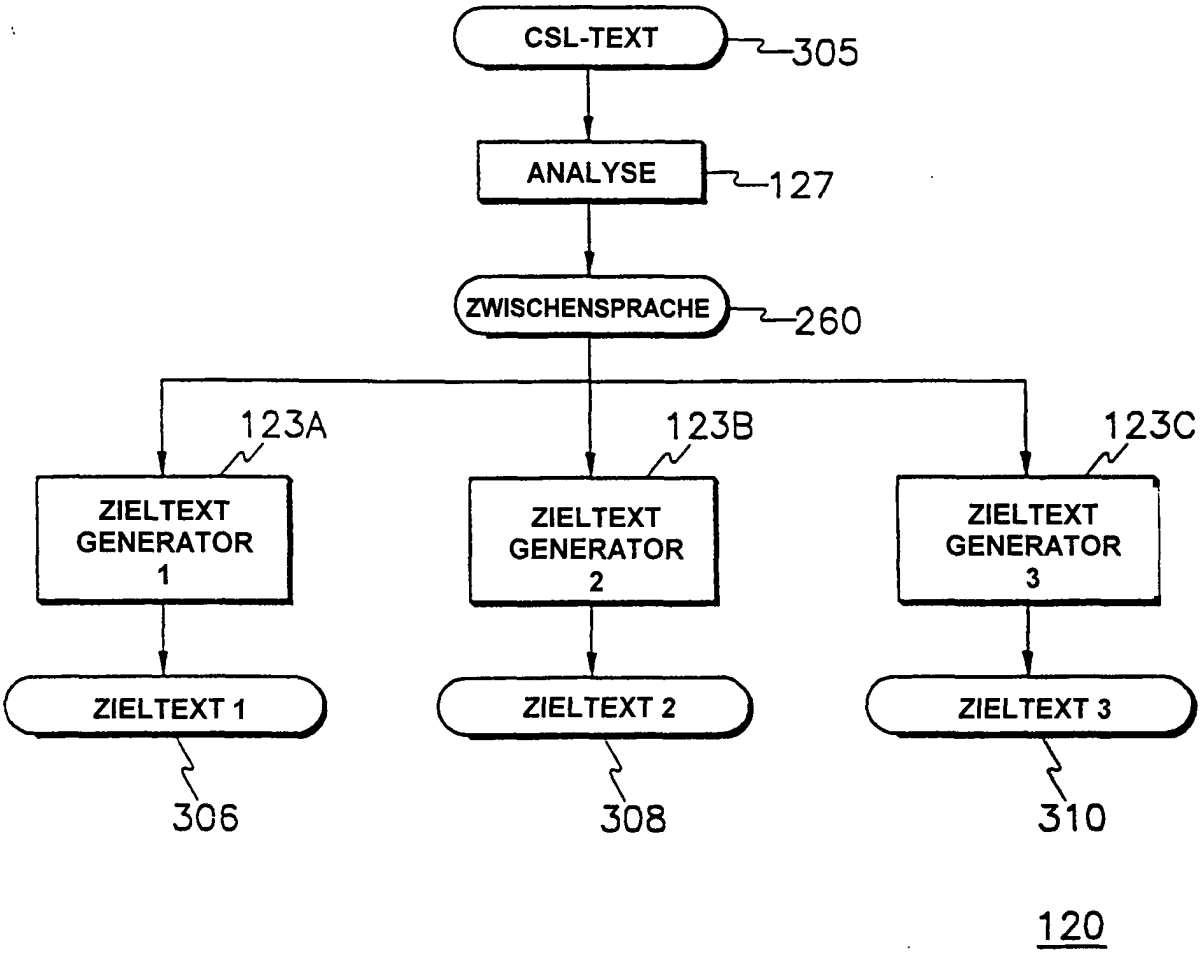
CSL-TEXT
305
ANALYSE
127
ZWISCHENSPRACHE
260
123A
123B
123C
ZIELTEXT GENERATOR 1
ZIELTEXT GENERATOR 2
ZIELTEXT GENERATOR 3
ZIELTEXT 1
ZIELTEXT 2
ZIELTEXT 3
306
308
310
120

Fig. 3

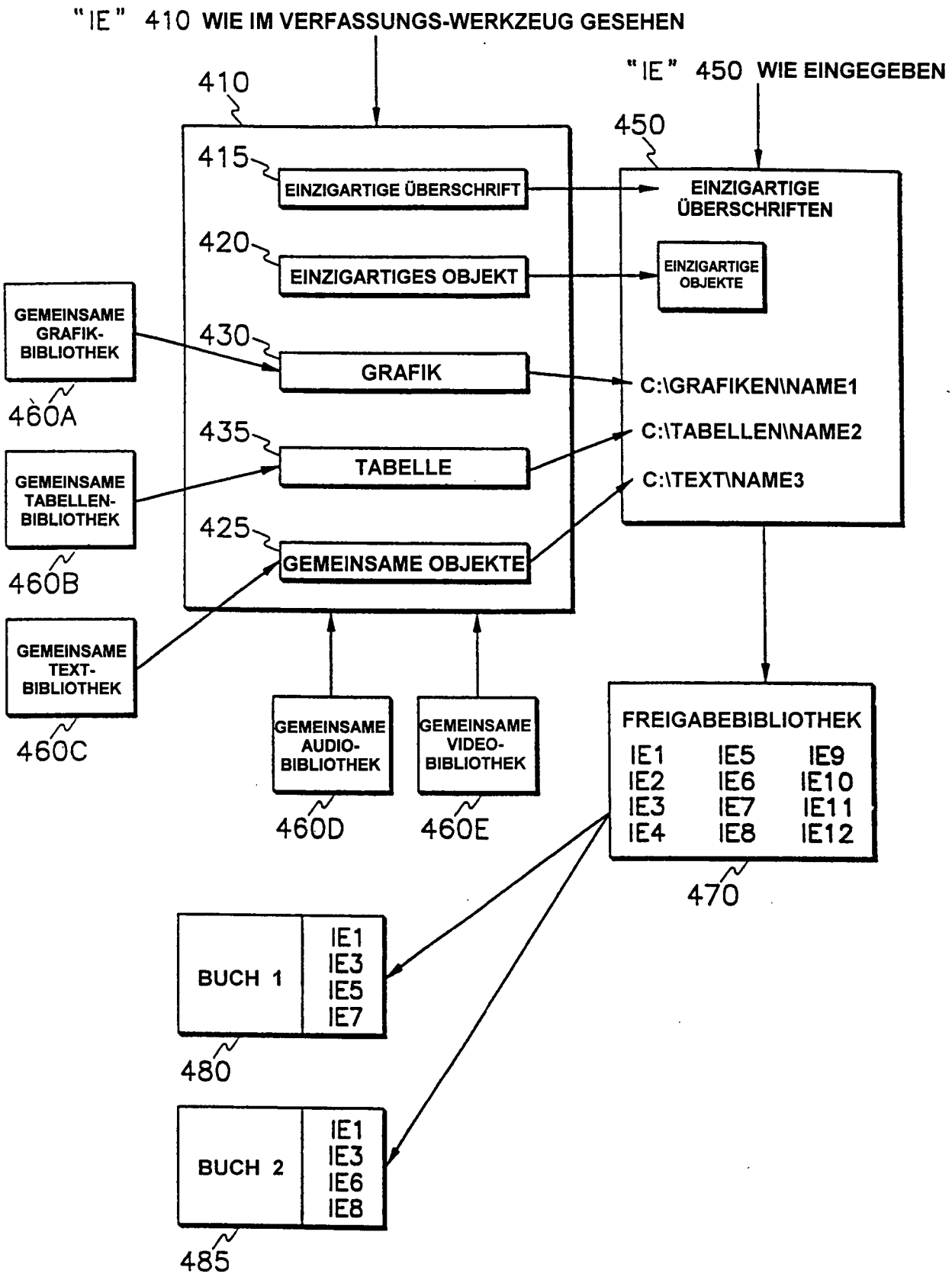
"IE" 410 WIE IM VERFASSUNGS-WERKZEUG GESEHEN
"IE" 450 WIE EINGEGEBEN
410
450
415
EINZIGARTIGE ÜBERSCHRIFT
EINZIGARTIGE ÜBERSCHRIFTEN
420
EINZIGARTIGES OBJEKT
EINZIGARTIGE OBJEKTE
GEMEINSAME GRAFIK-BIBLIOTHEK
460A
430
GRAFIK
C:\GRAFIKEN\NAME1
435
TABELLE
C:\TABELLEN\NAME2
GEMEINSAME TABELLEN-BIBLIOTHEK
460B
C:\TEXT\NAME3
425
GEMEINSAME OBJEKTE
GEMEINSAME TEXT-BIBLIOTHEK
460C
GEMEINSAME AUDIO-BIBLIOTHEK
460D
GEMEINSAME VIDEO-BIBLIOTHEK
460E
FREIGABEBIBLIOTHEK
IE1 IE5 IE9
IE2 IE6 IE10
IE3 IE7 IE11
IE4 IE8 IE12
470
BUCH 1
IE1 IE3 IE5 IE7
480
BUCH 2
IE1 IE3 IE6 IE8
485

Fig. 4

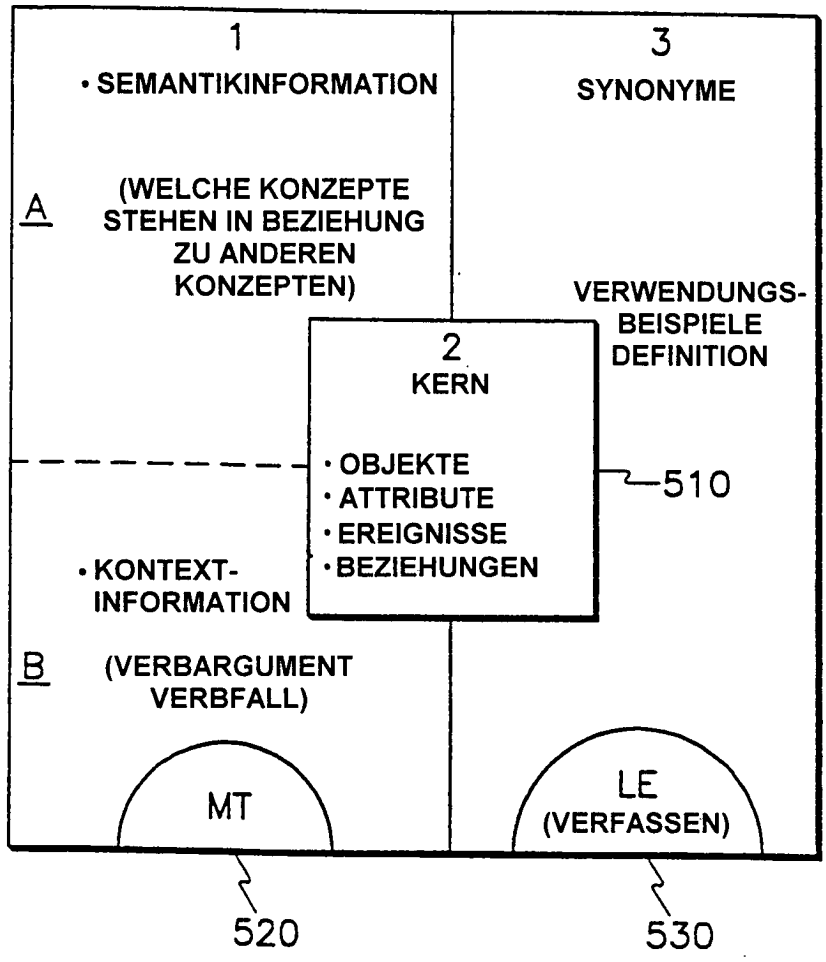
1
• SEMANTIKINFORMATION
A
(WELCHE KONZEPTE STEHEN IN BEZIEHUNG ZU ANDEREN KONZEPTEN)
3
SYNONYME
VERWENDUNGS-BEISPIELE DEFINITION
2
KERN
• OBJEKTE
• ATTRIBUTE
• EREIGNISSE
• BEZIEHUNGEN
510
• KONTEXT-INFORMATION
B
(VERBARGUMENT VERBFALL)
MT
LE
(VERFASSEN)
520
530
500

Fig. 5

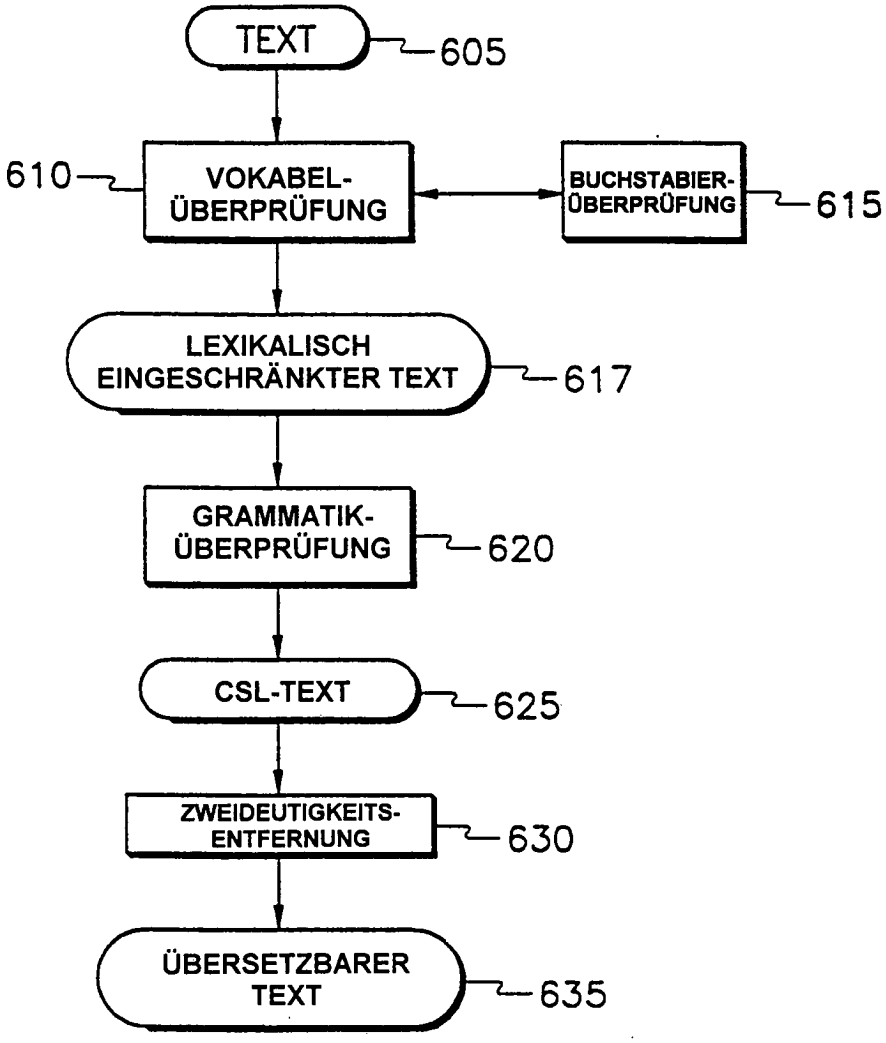
TEXT
605
610
VOKABEL-
ÜBERPRÜFUNG
BUCHSTABIER-
ÜBERPRÜFUNG
615
LEXIKALISCH
EINGESCHRÄNKTER TEXT
617
GRAMMATIK-
ÜBERPRÜFUNG
620
CSL-TEXT
625
ZWEIDEUTIGKEITS-
ENTFERNUNG
630
ÜBERSETZBARER
TEXT
635

Fig. 6

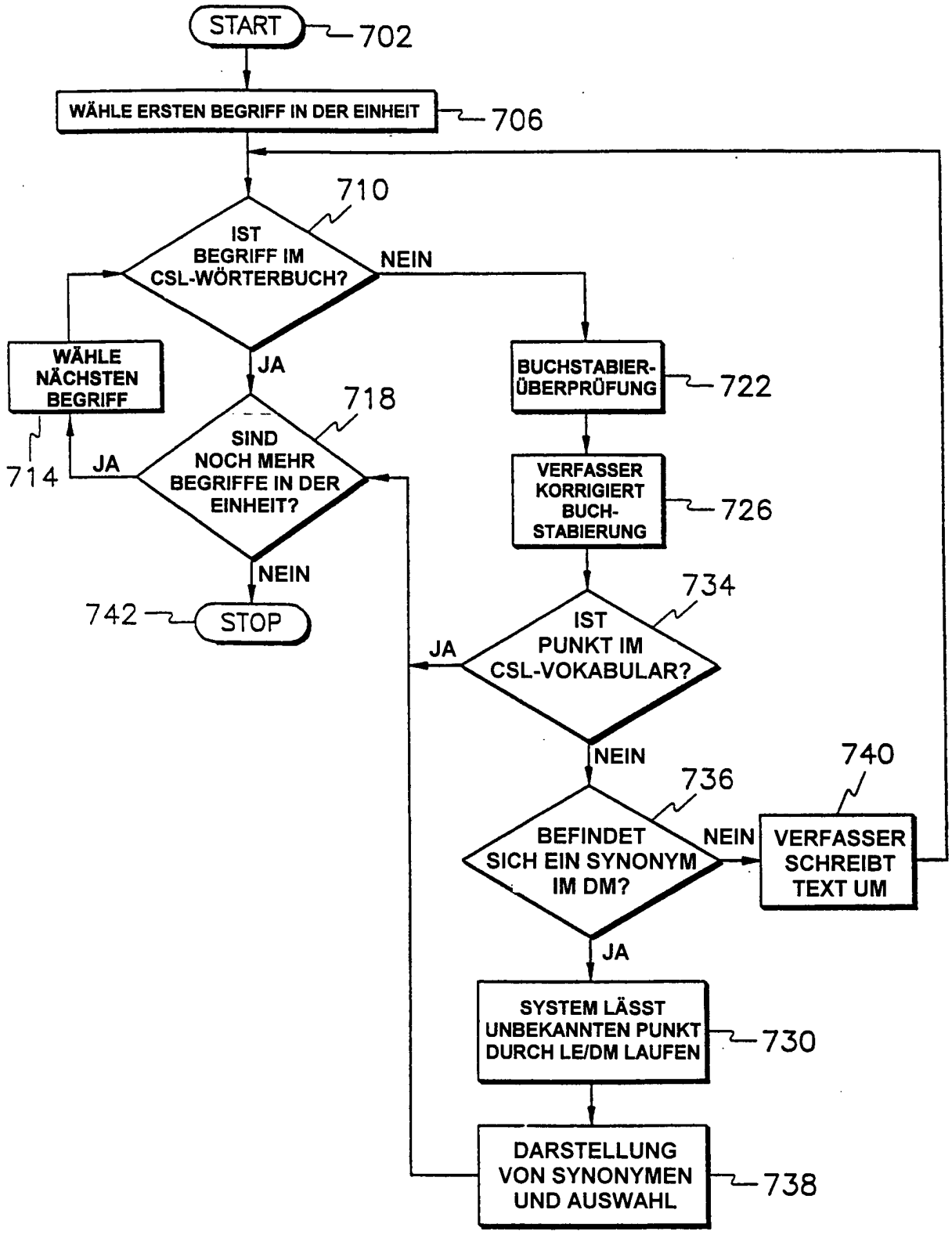
START
702
WÄHLE ERSTEN BEGRIFF IN DER EINHEIT
706
710
IST BEGRIFF IM CSL-WÖRTERBUCH?
NEIN
JA
WÄHLE NÄCHSTEN BEGRIFF
714
718
SIND NOCH MEHR BEGRIFFE IN DER EINHEIT?
JA
NEIN
742
STOP
BUCHSTABIER-ÜBERPRÜFUNG
722
VERFASSER KORRIGIERT BUCH-STABIERUNG
726
734
IST PUNKT IM CSL-VOKABULAR?
JA
NEIN
736
BEFINDET SICH EIN SYNONYM IM DM?
NEIN
740
VERFASSER SCHREIBT TEXT UM
JA
SYSTEM LÄSST UNBEKANNTEN PUNKT DURCH LE/DM LAUFEN
730
DARSTELLUNG VON SYNONYMEN UND AUSWAHL
738
700

Fig. 7

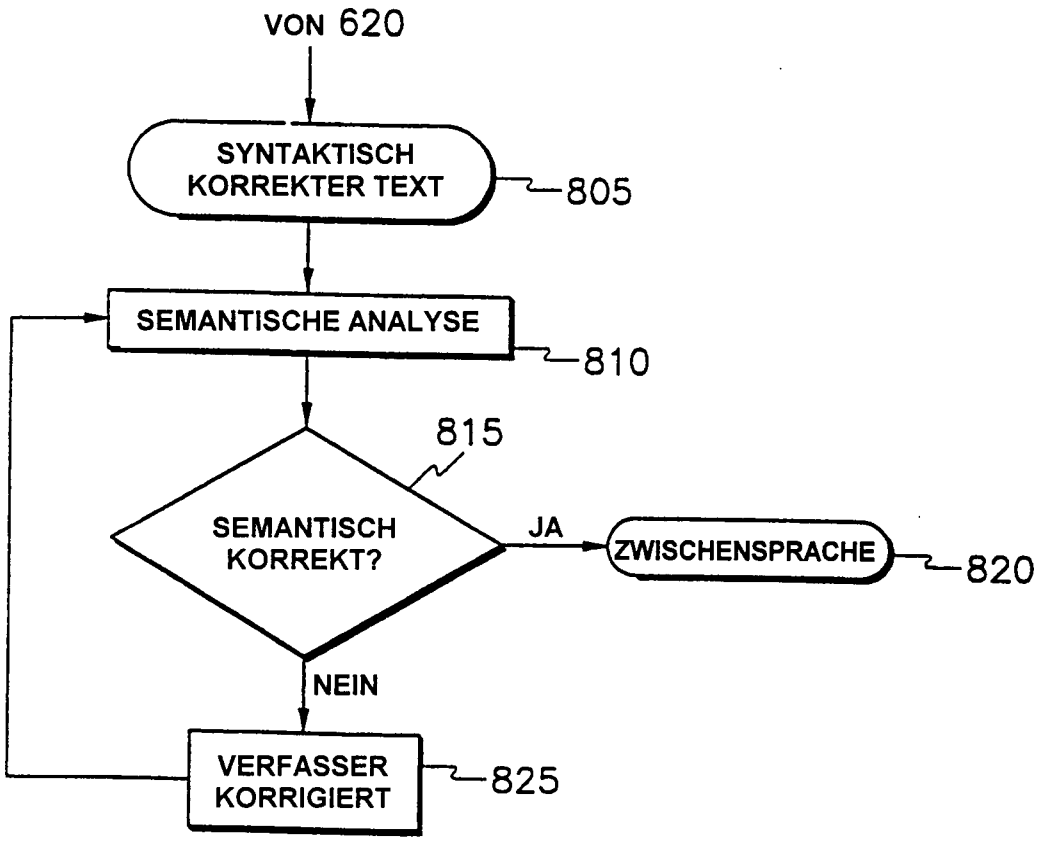
VON 620
SYNTAKTISCH KORREKTER TEXT
805
SEMANTISCHE ANALYSE
810
815
SEMANTISCH KORREKT?
JA
ZWISCHENSPRACHE
820
NEIN
VERFASSER KORRIGIERT
825

Fig. 8

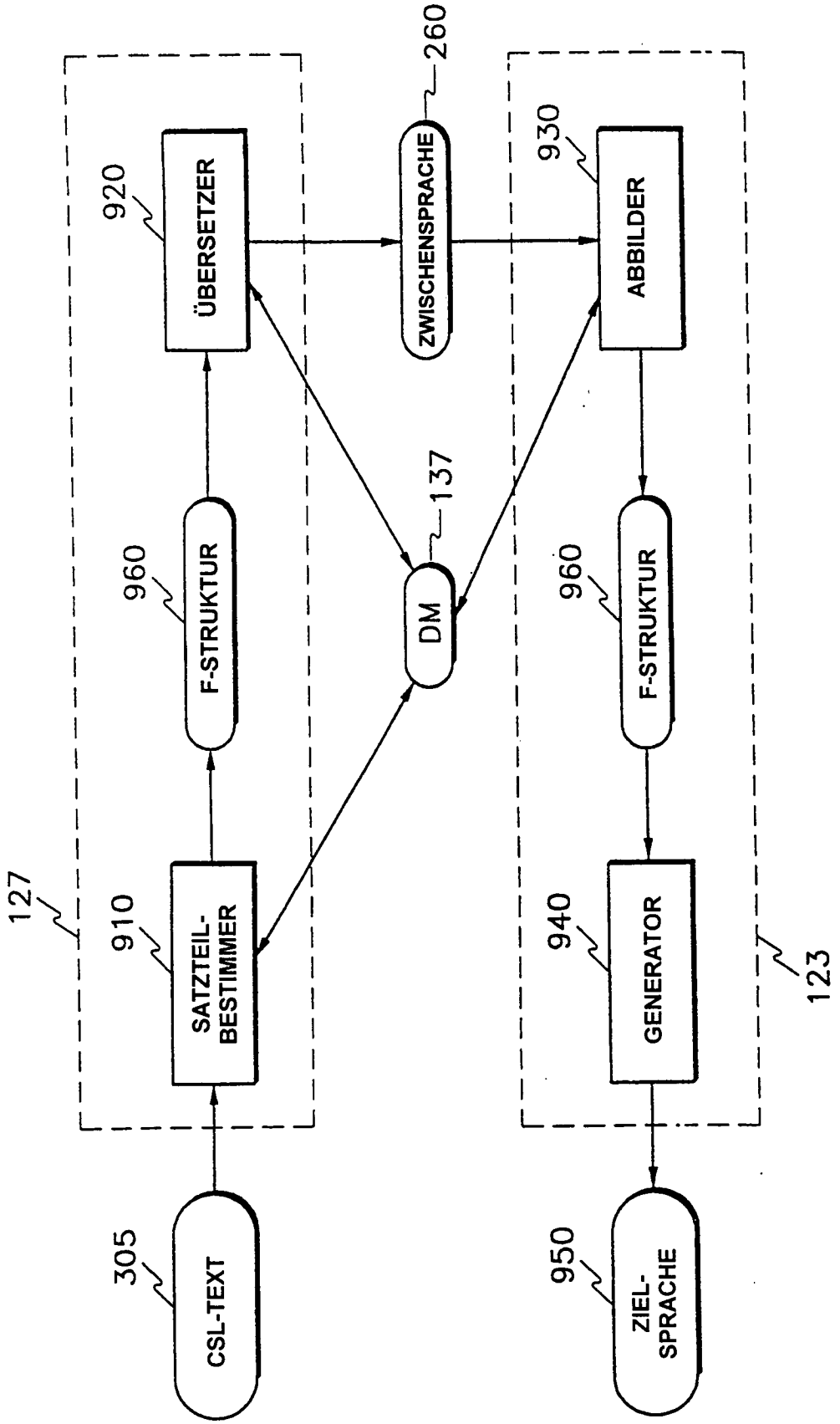
127
910
SATZTEIL-BESTIMMER
960
F-STRUKTUR
920
ÜBERSETZER
260
ZWISCHENSPRACHE
930
ABBILDER
960
F-STRUKTUR
940
GENERATOR
123
137
DM
305
CSL-TEXT
950
ZIEL-SPRACHE

Fig. 9